ON-CHIP OPTICAL INTERCONNECTS FOR CHIP MULTIPROCESSORS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nevin Kırman

February 2010

ON-CHIP OPTICAL INTERCONNECTS FOR CHIP MULTIPROCESSORS

Nevin Kırman, Ph.D.

Cornell University 2010

In this dissertation, we address the on-chip cross-core and -memory interconnection problem facing future large-scale chip multiprocessors (CMPs) through the use of silicon optical technology. CMOS-compatible silicon photonics is a disruptive technology that can potentially enable high-bandwidth, low-latency, and low-power interconnect solutions for both off- and on-chip data communication. Although the technology is still in its formative stages, and the more near-term application is chip-to-chip communication, rapid advances have been made in the development of on-chip optical interconnects and devices.

We first investigate the potential of optical technology to construct a low-latency, high-bandwidth shared bus supporting snoopy cache coherence in future CMPs. While not exhaustive, our initial investigation yields a hierarchical opto-electrical system that exploits the advantages of optical technology while abiding by projected limitations. Our evaluation shows that, compared to an aggressive all-electrical bus of similar power and area, significant performance improvements can be achieved using an opto-electrical bus. This performance improvement is largely dependent on the number of implemented wavelengths per waveguide.

We further improve on the data network. We present an all-optical approach to constructing data networks on chip that combines the following key features: (1) Wavelength-based routing, where the route followed by a packet depends solely on the wavelength of its carrier signal, and not on information either con-

tained in the packet or traveling along with it. (2) Oblivious routing, by which the wavelength (and thus the route) employed to connect a source-destination pair is invariant for that pair, and does not depend on ongoing transmissions by other nodes, thereby simplifying design and operation. And (3) passive optical wavelength routers, whose routing pattern is set at design time, which allows for area and power optimizations not generally available to solutions that use dynamic routing. We construct such an all-optical network and propose a connection-based operation. Our evaluation shows that our approach is competitive with prior proposals from the performance standpoint, yet it yields significantly more power-efficient designs.

BIOGRAPHICAL SKETCH

Nevin Kırman completed B.S. degrees in Electronics and Communication Engineering and Computer Engineering in 2002 and 2003, respectively, both from Istanbul Technical University (ITU), Turkey. She joined Cornell University in Fall 2003 to pursue M.S./Ph.D. degrees in Electrical and Computer Engineering. She earned her M.S. degree in 2007. Besides her main research focus on on-chip optical interconnects for chip multiprocessors, she did research in areas of checkpointed processor architectures, reconfigurable processor architectures, and memory system design for chip multiprocessors.

To my beloved family

ACKNOWLEDGEMENTS

First, I would like to sincerely thank Prof. José F. Martínez, my adviser, for continuously providing us with very crucial and effective feedback that have brought dramatical improvements, a lot of insights, and different perspectives to our research projects, writings, and presentations. His guidance, help, inspiration, and dedication throughout were invaluable and indispensable in successfully completing this work and the graduate study. Also, I am very grateful for the excellent research environment and opportunities he established for his group. I always consider myself very fortunate to have him as my adviser; his contributions to my academic and personal development are countless.

I am also greatly thankful to my sister Meyrem Kırman for always being with me, in good and in difficult times. Her help and support is always comforting, working together is always a pleasure.

Many thanks also to my committee members Prof. Rajit Manohar and Prof. Michal Lipson for their useful insights and feedback, to my other collaborators for the valuable teamwork, to the other members of my research group for insightful discussions, and to all my colleagues at Computer Systems Laboratory for providing such a friendly and stimulating research environment.

Finally, the support and encouragement of my family was endless, for which I am enormously grateful. They are continuous source of motivation and inspiration.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . x

1 Introduction* 1

2 Optical Technology Overview* 4
2.1 Transmitter . . . 4
2.2 Waveguide . . . 6
2.3 Receiver . . . 7

3 Opto-Electrical Bus Architecture* 9
3.1 CMP Architecture . . . 9
3.2 Optical Medium . . . 10
3.3 Bus Design . . . 13
3.3.1 Protocol . . . 13
3.3.2 Topology . . . 14

4 Opto-Electrical Bus Evaluation* 24
4.1 Experimental Setup . . . 24
4.1.1 Electrical Bus . . . 25
4.1.2 Opto-electrical Bus . . . 29
4.1.3 Applications . . . 29
4.1.4 Bandwidth Characterization . . . 30
4.2 Performance Evaluation . . . 30

5 All-optical Network Using Wavelength-based Oblivious Routing* 35
5.1 CMP Architecture . . . 37
5.2 Network Substrate . . . 38
5.2.1 Wavelength assignment . . . 39
5.2.2 Wavelength-path layout . . . 40
5.2.3 Wavelength-router design . . . 43
5.2.4 Transmitter/receiver interface . . . 44
5.2.5 Multiple network layers . . . 46
5.3 Network Operation . . . 48
5.3.1 Connection Protocol . . . 50
5.3.2 Network-Layer Selection . . . 52
5.3.3 Protocol Network Layers . . . 53
5.3.4 Hardware Support . . . 54
5.3.5 Optimizations . . . 56

6 Performance Evaluation of Optical Networks* 59
6.1 Experimental Setup 59
6.2 Applications 64
6.3 Performance Evaluation 65
6.4 Performance Analysis 66

7 Power Evaluation of Optical Networks* 73
7.1 On-chip Electrical Power Estimation 73
7.2 Optical Power Estimation 75
7.3 Results 84

8 Related Work* 88

9 Conclusions* 93

A Core Frequency Estimation* 94

B Wavelength Paths Found by the Genetic Algorithm 97

C Support for Credit-Based Control Flow* 100

D Design-Space Exploration of Xbar-Bcast Optical Networks 102

E Evaluation of a Circuit-Switched Hybrid Elecrical-Optical Network 106
E.1 Performance Evaluation 106
E.2 Power Evaluation 109

Bibliography 112

LIST OF TABLES

2.1 General characteristics of silicon and polymer waveguides 6

3.1 Delays of optical components at different technology nodes . . . 15
3.2 Area and power characterization of different optical bus topologies 17
3.3 Power losses incurred by various optical components and events 22

4.1 Optical bus evaluation: Processor core in the modeled CMP . . . 25
4.2 Optical bus evaluation: Memory subsystem in the modeled CMP 26
4.3 Area and power characterization of baseline electrical buses . . . 28
4.4 Application descriptions, simulated problem sizes, and observed global L2 miss rates 30
4.5 Parallel efficiencies of simulated SPLASH-2 applications 34

6.1 Processor core of the modeled system 60
6.2 Memory subsystem of the modeled system 61
6.3 Evaluated configurations of the proposed network 62
6.4 Applications' simulated problem sizes 65
6.5 Connection statistics in the proposed network 70
6.6 Fraction of all data supplies by sharer caches with existing connection in the proposed network 71

7.1 Electrical switches/(de)multiplexers in the evaluated optical networks 74
7.2 Component counts in the evaluated optical networks 75
7.3 Loss values for unit optical components/events used in optical power-loss analysis of the evaluated networks 76
7.4 Power consumption breakdown for the evaluated optical networks 84

A.1 ITRS parameters used to calculate the processor frequencies at different technology nodes 96

B.1 Full listing of the routing schemes found by the genetic algorithm - Part I 98
B.2 Full listing of the routing schemes found by the genetic algorithm - Part II 99

D.1 Electrical switches in the evaluated Xbar-Bcast optical networks . 103
D.2 Component counts in the evaluated Xbar-Bcast optical networks 103
D.3 Power consumption breakdown for the evaluated Xbar-Bcast optical networks 104

E.1 Electrical routers in the circuit-switched hybrid electrical-photonic network 109

E.2 Component counts in the circuit-switched hybrid electrical-photonic network 110
E.3 Power consumption breakdown of the circuit-switched hybrid electrical-photonic network 111

LIST OF FIGURES

2.1 Main components in on-chip optical transmission 5

3.1 Simplified CMP floorplan and high-level system organization . . 15

4.1 Modeled electrical baseline address and data networks 27
4.2 Applications' bandwidth demand characterization 31
4.3 Performance improvements of four-node opto-electrical buses . . 32
4.4 Latency breakdown of bus transactions in evaluated buses 33

5.1 Example optimal wavelength assignment by Aggarwal et al. . . 39
5.2 Wavelength assignment examples 41
5.3 The 6x4 two-dimensional torus adopted in this study 42
5.4 Passive wavelength-router implementation 43
5.5 Simplified diagram of transmitter and receiver interfaces at end nodes 45
5.6 Concentric layout of multiple network layers 46
5.7 Circular layout of a torus network layer 47
5.8 Protocol state diagrams for Rx-side and Tx-side 49
5.9 Node's interface to data and protocol network layers 55

6.1 Performance of the optical networks 65
6.2 Average latency breakdown in the address and data networks . . 66
6.3 Breakdown of data-transmission requests in the proposed network 67
6.4 Breakdown of connection-lookahead requests in the proposed network 68
6.5 Study on effectiveness of the proposed optimizations in the optical network 72

7.1 Light-path model of Xbar-Bcast 78
7.2 Light-path model of Xbar-Arb 79
7.3 Wavelength-based light-power distribution in Oblivious network 81
7.4 Followed rules in optical loss estimation 83
7.5 Sensitivity of Xbar-Arb's optical power to optical loss parameters 85
7.6 Sensitivity of Oblivious's optical power to optical loss parameters 86

A.1 Leakage power projections 95

C.1 Credit-flow timing diagram 100

D.1 Performance comparison of the explored Xbar-Bcast networks. . 104

E.1 Layout of the evaluated circuit-switched hybrid electrical-photonic network 107
E.2 Approximated performance of the circuit-switched hybrid electrical-photonic network 108

CHAPTER 1

INTRODUCTION*

Current research and technology trends indicate that future chip multiprocessors (CMPs) may comprise tens or even hundreds of processing elements. An important hurdle towards attaining this scale, however, is the need to feed data to such large numbers of on-chip cores. This can only be achieved if architecture and technology developments provide sufficient chip-to-chip and on-chip communication performance to these large-scale CMPs.

Optical technology [24, 42, 86] and 3D integration [61, 66] are two potential solutions to current and projected limitations in *chip-to-chip* communication performance. Still, *on-chip* communication faces considerable technological and architectural challenges of its own. On the one hand, global on-chip interconnects do not scale well with technology [33, 35]. Although delay-optimized repeater insertion [4, 33, 63] and proper wire sizing [32] can keep the delay nearly constant, this comes at the expense of power [33, 38] and active area, as well as a reduction in wire count (and thus bandwidth). Techniques for optimizing the power-delay product have been developed [5, 38], but unfortunately their most obvious shortcoming is that neither power nor latency are optimal. This, combined with various other technological issues such as manufacturability, conductivity, crosstalk, etc., constitute important roadblocks for future electrical interconnects [35]. On the other hand, electrical on-chip network designs are

*

likely to be severely constrained by the limited on-chip power budget, as well as long multi-hop latencies.

CMOS-compatible silicon photonics is a disruptive technology that can potentially enable high-bandwidth, low-latency, and low-power interconnect solutions for both off- and on-chip data communication. Whereas ten years ago the electrical-optical translation costs and CMOS incompatibility were viewed as insurmountable barriers for the use of optics in on-chip communication, today the outlook is dramatically more optimistic. Due to rapid progress in the past five years in CMOS-compatible detectors [77], modulators [3], and even light sources [69], the latest ITRS considers on-chip optical interconnects as a potential replacement for global wires by 2013 [35]. In global signaling applications, optical interconnects have the potential to fare favorably against their electrical counterparts due to their high speed, high bandwidth, low on-chip power, good electrical isolation, low electromagnetic interference, and other benefits [52]. Although the technology is still in its formative stages, there is now enough understanding and data regarding on-chip, CMOS-compatible, optical components to consider the broader architectural trade-offs in designing an on-chip optical network for future high performance microprocessors.

In the first part of this dissertation, we investigate the potential of optical technology as a low-latency, high-bandwidth shared bus supporting snoopy cache coherence in future CMPs. We discuss possible optical bus organizations in terms of power, scalability, architectural advantages, and other implementation issues, as well as the implications on the coherence protocol. Through a carefully projected case study for a 32nm CMP, we conduct the first evaluation of on-chip optical buses for this application. This initial step yields insights

into the advantages and current limitations of the technology to catalyze future interdisciplinary work.

In the second part of the dissertation, we improve on the optical data network. An all-optical approach to constructing an on-chip data network constitutes a very attractive proposition from the performance standpoint, and a careful design can deliver a fundamentally power-efficient solution that is reasonably robust to technology considerations. We argue for such an approach. Specifically, our proposed optical-routing based solution is an all-optical network that routes optical signals based on wavelength information, eliminating the need for intermediate O-E/E-O conversions. Oblivious routing is facilitated through passive optical routers, simplifying design and operation and allowing for area and power optimizations not generally available to solutions that use dynamic routing.

CHAPTER 2

OPTICAL TECHNOLOGY OVERVIEW*

We consider on-chip modulator-based optical transmission (Figure 2.1), which comprises three major components: transmitter, waveguide, and receiver. We briefly describe each component.

2.1 Transmitter

Optical transmission requires a laser source, a modulator, and a modulator driver (electrical) circuit. The laser source provides light to the modulator, which transduces electrical information (supplied by the modulator driver) into a modulated optical signal.

While both off- and on-chip laser sources are feasible, in this work we opt for an off-chip laser source because of its greater on-chip power, area, and cost savings. As the light enters the chip, optical splitters and waveguides (not shown in Figure 2.1) route it to the different modulators used for actual data transmission. These distribution paths are a source of signal losses.

The modulator translates the modulator driver's electrical information into a modulated optical signal. High-speed electro-optic modulators are designed such that injection of an electrical signal changes the refractive index or the absorption coefficient of an optical path. Among different types of proposed modulators [6, 43, 45, 65], the most recent optical resonator-based implementations

*

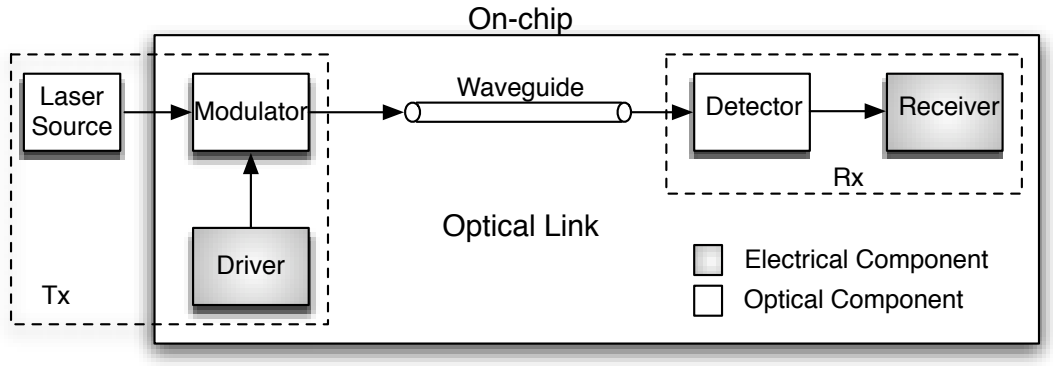


Figure 2.1: Simplified diagram showing the main components involved in on-chip optical transmission. Tx and Rx stand for transmitter and receiver, respectively.

are preferable for integrated circuit design, due to their low operating voltage and compact size [6]. We assume this type of modulator in our work.

Modulators are the optical equivalent of electrical switches (or transistors acting as such). Their performance in part is dependent on the on-to-off light intensity ratio, called the *extinction ratio*, which is dependent upon the strength of the electrical input signal. Higher extinction ratio is better for proper signal detection. A poor one may cause transmission errors in the channel. This ratio also puts constraints on the number of transmitters that can time-share the same wavelength on the same channel. An extinction ratio greater than 10dB has been recently reported with high input signal swing [3].

Modulator size is another important criterion for integrated applications. There has been significant recent activity towards realizing compact-sized modulators. Already 10μm ring-modulators (circularly shaped) have been proposed [3], and their size is likely to be reduced with each successive generation,

albeit bounded by lithographic process and bending curvature limitations.

The modulator driver consists of a series of inverter stages driving the modulator's capacitive load. A smaller capacitance will improve the power and latency specifications of the overall transmitter, thereby requiring fewer stages.

2.2 Waveguide

Waveguides are the paths through which light is routed. The refractive index of the waveguide material has a significant impact on optical interconnect bandwidth, latency, and area. For on-chip applications, silicon (Si) and polymer are the most promising materials. Some of the most relevant features of silicon and polymer waveguides are given in Table 2.1.

Table 2.1: General characteristics of silicon and polymer waveguides.

Waveguide	**Si**	**Polymer**
Refractive index	3.5	1.5
Width (μm)	0.5	5
Separation (μm)	5	20
Pitch (μm)	5.5	25
Time of flight (ps/mm)	10.45	4.93

The smaller refractive index of polymer waveguides results in higher propagation speed. On the other hand, polymer waveguides require a larger pitch than Si, which reduces bandwidth density (the number of bits that can be transmitted per unit surface area).

For integrated applications, an additional disadvantage of polymer waveguides is the lack of a compact modulator. Although modulators exist for both silicon [3] and polymer waveguides [60], polymer-based modulators are bulky,

and require high voltage drive for high frequency operation. These drawbacks limit their applicability to on-chip optical links.

Polymer waveguides are feasible in a transmission system based on VCSELs (Vertical Cavity Surface Emitting Laser) [69], where the modulator is not required. However, a VCSEL-based solution tends to increase on-chip power with the added complexity of on-chip/flip-bonded laser sources. Also, the light is emitted vertically and must be transferred to the horizontal chip surface, which requires integrated mirrors and sophisticated lithographic technologies. For these various reasons, we choose to study systems using silicon waveguides, although we understand that with technological advances feasible options might become available with polymer waveguides.

2.3 Receiver

An optical receiver performs the optical-to-electrical conversion of the light signal. It comprises a photodetector and a trans-impedance amplifier (TIA) stage. In wave division multiplexing (WDM) applications, which involve simultaneous transmission at different wavelengths per waveguide, the receiver also requires a wave-selective filter for each received wavelength.

The photodetector that is most often proposed is a P-I-N diode [85]. The photodetector's *quantum efficiency* is an important figure of merit for the system. A high quantum efficiency means lower losses when converting optical information into electrical form. Detector size is also an important criteria for both compactness and next stage capacitance. Typically, the detector has large base capacitance and pose a design challenge for high-speed gain stages following it.

The TIA stage converts photodetector current to a voltage which is thresholded by subsequent stages to digital levels [59]. To achieve high-gain and high-speed detection, an analog supply voltage higher than the digital supply voltage may be required, thereby requiring higher power.

CHAPTER 3

OPTO-ELECTRICAL BUS ARCHITECTURE*

In this chapter, we explore the opportunities and challenges of building an optical bus for a particular application and technology node. Working bottom-up, we first determine a reasonable CMP organization (in terms of cores, memory hierarchy, operating frequency, etc.), using available data from ITRS and other sources. Then, we address the specifics of designing a cache-coherent network with integrated optical system components (Section 3.3). In the following chapter, we evaluate the optical bus designs.

3.1 CMP Architecture

We target a 32nm process technology, and assume a 400mm^2 die area. Assuming 10mm^2 per core+L1 at 65nm (for a stripped version of an out-of-order Power4-like core [41]), and extrapolating to 32nm, we find that 64 cores fit comfortably on the die (occupying 40% of the die area), with enough additional space to allocate L2 caches (20%), interconnect (15%), controllers for off-chip L3 cache and memory, and other system components (25%). The area breakdown closely follows the one in [26].

We opt for sixteen L2 caches, each shared among four cores, as a compromise between the demonstrated benefits of cache sharing [25, 34, 70] and the area/power overhead [41]. Using CACTI4.1 [68], we find sixteen 2MB L2 caches to fit in the allocated area.

*

We reasonably assume that the use of chip-to-chip optical technology will precede its on-chip application [9], and set off-chip pin bandwidth to 256GB/s and 128GB/s to L3 and memory, respectively. The aggregate pin bandwidth is therefore 384GB/s (3Tbit/s), which is well within current industry projections for our proposal's time frame [24, 42].

We estimate that core frequency will remain approximately constant in subsequent technologies, in agreement with [13]. (For a quantitative analysis, see Appendix A.) Thus, if we reasonably assume a 4GHz core frequency at 65nm, we can set core frequency in our 32nm CMP also to 4GHz.

3.2 Optical Medium

Optical waveguides do not lend themselves gracefully to H-tree or highly angled structures that may be more common in electrical topologies, for turns and waveguide crossings may result in significant signal degradation. This is aggravated when attempting to lay out multiple waveguides for multi-bit transmission, which is the case in a typical bus. Instead, we propose to build upon a simple loop-like structure, which is much better suited to the structural characteristics of optical waveguides. In the rest of this section, we discuss the design implications of this structural choice.

The proposed loop-shaped bus comprises optical waveguides (residing on a dedicated Si layer) that encircle a large portion of the chip (Figure 3.1). Multiple nodes connected to the bus, each of them responsible for issuing transactions on behalf of a processor or a set of processors, are equipped with necessary transmitters and receivers to interface with the optical medium, as explained

earlier (Chapter 2).

We assume a bus comprising a total of b address, data, and snoop response bits (and thus waveguides). We further presume the availability of w wavelengths per waveguide through wave division multiplexing (WDM) [19, 40], which we use to realize a w-way multibus.

We explore two typical ways to multiplex this multibus organization: by address and by node. In multiplex by address, where wavelengths are assigned to different address spaces, any node can drive any of the w wavelengths, and thus requires arbitration. On the other hand, multiplexing by node gives each of the n nodes exclusive access to $\frac{w}{n}$ wavelengths (with w an integer multiple of n), which we will see has numerous advantages; however, the downside is that the number of nodes directly connected to the bus is then limited to w at best. (Other options are possible, for example leveraging WDM to decrease the number of physical waveguides by w. For the sake of simplicity, we leave this and other options for future work.)

An important consideration for both organizations is to prevent the light from circulating around the loop for more than one complete cycle, or older messages can cause undesirable interference. This can be easily handled in multiplex-by-node organizations by placing attenuator immediately before each modulator, to act as "sink" for the corresponding wavelength once the signal goes full circle. Alternatively, both multiplex-by-address and multiplex-by-node organizations may use an attenuator to "open" the loop, as long as modulators transmit in both directions simultaneously.

One power advantage of the multiplex-by-node organization is that it only

requires $nb\frac{w}{n}$ transmitters ($\propto n$ if $w = n$), vs. nbw transmitters ($\propto n^2$ if $w = n$) in the multiplex-by-address organization. Since the optical power in a continuous laser source based system is dependent upon the number of modulators (Section 3.3.2), this difference may result in substantial optical power advantage for the multiplex-by-node organization.

Another power advantage of multiplex-by-node over multiplex-by-address is the possibility to optimize light power through individual coupling-ratio tuning at detectors at design time. This is because in multiplex-by-node organizations, the relative position of each detector with respect to the (sole) transmitter is known for every wavelength, and thus coupling at each detector can be designed to absorb just the right fraction of light power as to allow for efficient delivery to all detectors involved. In multiplex-by-address organizations, coupling at all detectors must be identical, since the signal may come from any one of the transmitters on the same wavelength, and thus the relative order in which they tap onto the signal is not known at design time. It can be shown mathematically that this results in wasted light power.

A third source of power waste in the multiplex-by-address organization comes from the fact that modulators do leak some light into the waveguide even in the "off" position. The more modulators coupled to a particular wavelength, the more aggregate light power leaks into the waveguide. In order for detectors to identify "on" and "off" states correctly, a proportional current bias must be applied to the receivers, which may result in a significant power waste.

For all the above reasons, in this work we opt for the more practical multiplex-by-node organization.

3.3 Bus Design

We propose an opto-electrical hierarchical bus, where the optical loop constitutes the top level of the hierarchy, and nodes deliver information to processors via electrical sublevels. Figure 3.1 depicts a possible four-node organization for our 64-processor CMP, where each node is shared among four electrically interconnected L2 caches.

Our bus comprises an address/command bus, a data bus, and a snoop response bus. We allocate 64 bits to address/command (including ECC and tag bits), 72 bits to data (including 8-bit ECC and assuming that tags are provided at the header), and 8 bits per snoop response. Therefore, the number of waveguides is 136 for address/command plus data buses, and $8n$ to support snoop responses (each node provides w snoop responses using $\frac{w}{n}$ different wavelengths, for a total of $\frac{8w}{\frac{w}{n}} = 8n$ waveguides).

3.3.1 Protocol

Before delving into the details of a design space exploration, we give a high-level description of the bus protocol. The specifics of the cache coherence protocol are not relevant here; we focus on the handling of coherence requests by the split-transaction, fully pipelined hierarchical bus.

L2 cache accesses by processors may result in coherence requests, which travel down the electrical sublevel to the corresponding node where they are enqueued. Node switches arbitrate among the incoming coherence requests, and broadcast the winner(s) on the optical address bus.

Every node snoops in the requests put on the optical address bus by every other node. (Recall that each node transmits through different wavelengths.) Then, nodes arbitrate among concurrent requests, using the same finite state machine so that they all reach the same outcome independently. (This requires factoring in requests even at their originating switch.) Next, the selected requests are delivered to all caches simultaneously, and the rest are retried later. Caches compose individual snoop responses, which are relayed back down to the optical snoop response bus, which again all nodes read and process concurrently. Finally, the appropriate decision is made and the final snoop result is propagated up to the caches where the appropriate action is taken. Eventually, if indicated, data is generally sent down to the optical data bus (after winning arbitration over possibly competing responses from other caches in the same node), which the original requesting node collects and sends up to the requesting L2 cache.

3.3.2 Topology

Different literature sources offer varying projections on the number of available wavelengths per on-chip waveguide. Chen et al. [19] project that the number of wavelengths per waveguide will increase by four with each technology generation, reaching thirteen wavelengths at 32nm, while Kobrinsky et al. [40] assume an increase of one wavelength every other generation, resulting in three-four wavelengths at 32nm. Accordingly, we explore a range of four to twelve available wavelengths per waveguide.

We investigate several possible bus topologies, deriving for each of them

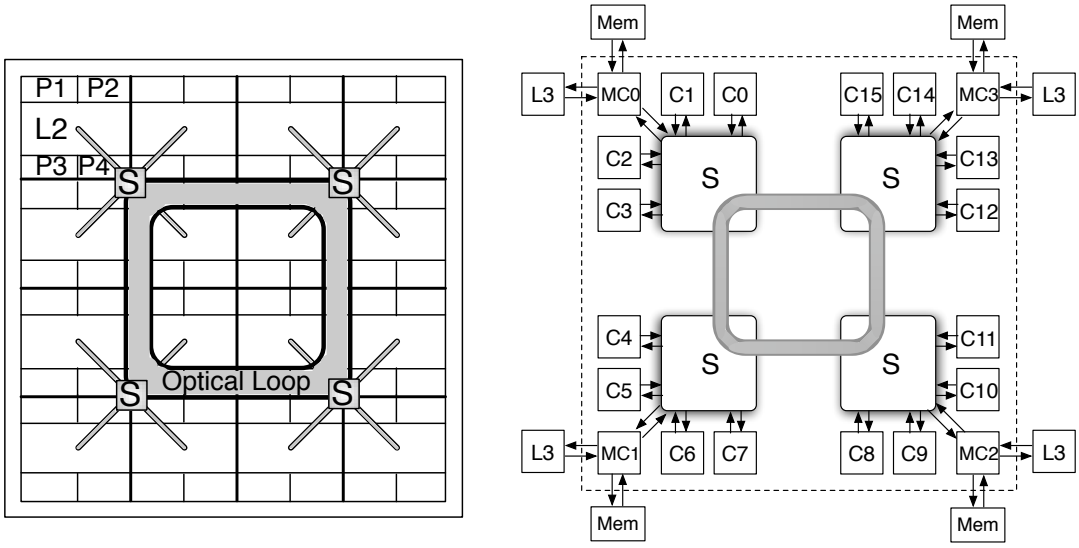


Figure 3.1: Simplified CMP floorplan diagram (left) and high-level system organization (right), showing the optical loop and the rest of the hierarchical bus. In the figures, S, MC(0-3), and C(0-15) stand for switch (separate switches for address/snoop and data buses), memory controller, and (L2) cache, respectively.

Table 3.1: Delays of optical components at different technology nodes [19].

Delays of Optical Components [19]			
Technology	**45nm**	**32nm**	**22nm**
Modulator driver (ps)	25.8	16.3	9.5
Modulator (ps)	30.4	20	14.3
Detector (ps)	0.6	0.5	0.4
Amplifier (ps)	10.4	6.9	4.0
Si waveguide delay (ps/mm)	10.45	10.45	10.45

area and power. Table 3.2 lists such topologies. In the table, H-*n*x*k*A*k*D (H for Hierarchical) designates a topology with *n* nodes on the optical bus and *k* address (data) wavelengths per node, totaling to *nk* wavelengths per waveguide in the address (data) bus. Beyond the optical loop, appropriately-sized electrical switches connect each of the sixteen quad-processor nodes to the network (hence the name Hierarchical). We sweep through all possible configurations

given the WDM projections stated earlier on in this section: $k \in \{1, 2, 3\}$ for $n = 4$, and $k = 1$ for $n = 8$. For the sake of completeness, we also investigate a F-16x1A1D (F for Flat) topology, which requires no electrical routers (hence the name Flat), but that is unrealizable under WDM projections.

In the case of four nodes and $k > 1$, we also investigate topologies with a more limited support for new address transactions per cycle, H-nx1AkD, as we empirically observe in the course of our evaluation (Chapter 4) that this is enough to satisfy the applications' bandwidth demand on the address bus in the simulated system under consideration. This should generally result in area and power savings. Similarly, for the sake of area and power savings in the case of eight and sixteen nodes, we explore reducing the electrical snoop bandwidth to four (matching the bandwidth of the H-4x1AkD topologies). This is indicated by appending (4S) to the topology encoding.

Frequency Estimation

We estimate the operating frequency of the bus by calculating the time needed for the light to travel from any node to the farthest node on the (unidirectional) optical loop, so that information can be transmitted to all nodes in one bus cycle. With the loop bus centered on the die (Figure 3.1), and through simple geometric calculations, we estimate its total length to be 36mm, 45mm, and 45mm for 4-, 8-, and 16-node topologies, respectively. If we assume for simplicity that all neighboring nodes are equidistant, then the distance between any two nodes that are farthest apart is 27mm, 39.4mm, 42.2mm, respectively. Using the waveguide and optical-component delays provided in Table 3.1, and accounting for 4 FO4 latching delay (estimated using ITRS data), we obtain the maximum operating

Table 3.2: Area and power characterization of different optical bus topologies. Tx/Rx stands for transmitter/receiver; α is switching activity factor. Total on-chip power is the sum of switch, wiring, Tx/Rx, and half the optical power components (due to a 3dB coupling loss (Section 3.3.2), only half of the optical power is actually consumed on chip). All dynamic power components in switching, wiring, and Tx/Rx columns assume α=1. For α=0.5, only the total sum is provided.

Optical Topology	**Snoop Requests /Bus clk**	**Area (mm^2)**				**Power (W)**					
		Active Si Layer		Metal Layer	Optical Layer	Electrical Level		Optical Level		Total On-chip	
		Switch	Tx/Rx			Switch	Wiring	Tx/Rx	Optical	(α=1)	(α=0.5)
H-4x1A1D	4	1.71	0.39	15.21	33.68	1.75	12.82	0.60	0.79	15.56	9.04
H-4x2A2D	8	2.72	0.78	24.42	34.10	3.03	20.59	1.19	1.58	25.60	15.13
H-4x3A3D	12	4.00	1.17	33.64	34.51	4.64	28.36	1.79	2.37	35.98	21.49
H-4x1A2D	4	1.93	0.56	15.21	33.86	2.06	12.82	0.85	1.13	16.30	9.73
H-4x1A3D	4	2.13	0.72	15.21	34.04	2.37	12.82	1.11	1.47	17.03	10.41
H-8x1A1D	8	4.05	1.89	12.21	51.64	4.50	10.30	3.07	6.35	21.05	15.44
H-8x1A1D(4S)	4	3.08	1.59	7.6	51.3	3.25	6.41	2.58	5.33	14.91	11.34
F-16x1A1D	16	14.38	10.05	N/A	77.08	16.70	N/A	16.78	39.06	53.01	50.90
F-16x1A1D(4S)	4	6.77	6.4	N/A	72.81	7.42	N/A	10.68	24.86	30.53	29.53

frequencies: 2.9GHz, 2.1GHz, and 2GHz, respectively. This implies that all three buses can run safely at 2GHz–exactly half the cores' frequency. (For simplicity, we assume that the electrical routers in the Hierarchical topologies can operate at this frequency regardless of their size.)

Area Estimation

We estimate the required areas on the active, optical, and metal layers for each organization (Table 3.2). All address, snoop, and data buses are considered in the area calculations.

In the active area, we account for electrical switches in each node, as well as transmitters and receivers on the optical bus. For simplicity, however, we do not include the area occupied by the repeaters in the electrical wiring, although we do include their contribution to power consumption later in this section. We use Orion [74] to estimate the area of input and output buffers, as well as the crossbar areas inside the switches. We assume four-entry input buffers to receive requests/addresses from each L2 cache, and single-entry input buffers for snoop request/response networks. In the data network, we allocate sixteen-entry buffers to collect data from each L2 cache, but compensate input buffer size at the optical end with optical width as follows: sixteen-, eight-, or four-entry input buffers in four (4x1D), eight (4x2D and 8x1D), or wider (4x3D and 16x1D) optical bus topologies, respectively. Output buffers are single-entry in all cases. We carefully specify the number of input and output ports considering the components connected to each switch (Figure 3.1), which in turn determines the number of input and output buffers, as well as the size of the crossbar in each case.

We estimate the active area taken up by transmitters and receivers required for the optical buses by conservatively assuming that modulator driver and TIA each occupy $50\mu m^2$, although standard scaling rules predict smaller areas for these components [55]. We assume $80\mu m^2$ modulators (10μm-diameter ring), $10\mu m \times 10\mu m$ detectors [19], and $80\mu m^2$ wave-selective filter areas (10μm-diameter ring resonator). Modulators and detectors consume area in both the active and optical layers; modulator drivers and TIAs are on the active layer, and filters are on the optical layer.

For the multiplex-by-node optical buses, the number of transmitters in each node is $\text{tx}_{\text{node}} = b_a a + b_d d + b_s s$, where b_a, b_d, and b_s are the number of address, data, and snoop-response bits, respectively, and a, d, and s are address, data, and snoop bandwidth per node, respectively. Since each node has to be able to receive all the transmitted information by other nodes, the total number of receivers is $(n-1)\text{tx}$, where n is the number of nodes on the optical bus, and $\text{tx} = n \cdot \text{tx}_{\text{node}}$ is the total number of transmitters on the bus. Therefore, while the number of transmitters is O(n), the number of receivers is $O(n^2)$.

The area occupied in the optical layer is calculated as the sum of waveguide, modulator, detector, and wave-selective filter areas. We assume the component areas specified above, and Si waveguide pitch as provided in Table 2.1.

The resulting active area is relatively modest, and the required optical layer easily fits within $400mm^2$ (Table 3.2).

Finally, we estimate the metal wiring area required for the electrical sub-interconnects in hierarchical organizations. We assume a global wire pitch of 400nm and wire length of 4.5mm and 2.25mm (estimated according to the floor-

plan in Figure 3.1) for four- and eight-node configurations, respectively. From each cache to its node, the links include single address and data paths, and as many snoop-response paths as needed in each topology (number of snoop requests per cycle in Table 3.2). From each node to a cache, the links include single data path and as many snoop-request and snoop-result paths as indicated in the table.

Power Estimation

We categorize the power consumption of the interconnect system into two: the power consumed in the electrical sublevels (switches and wiring), and the power consumed in the optical bus. Table 3.2 shows a detailed breakdown of power consumption in all topologies under consideration. We report power calculations for each component assuming full switching activity ($\alpha = 1$), but report total power consumption at full, as well as 50% activity ($\alpha = 0.5$).

We estimate the static and dynamic power consumed by the switches in the nodes again using Orion [74] following the structural assumptions outlined in Section 3.3.2.

The static and dynamic power consumption of the wires is estimated following the methodology in [32, 33] for power-delay optimized repeater insertion and wire sizing.[1] We estimate according to ITRS [35] projections that a minimum-sized repeater has approximately 1μW of leakage power consumption.

There are two main power components due to the optical loop: electrical

[1] We estimate 26ps/mm repeatered wire delay.

and optical power. Electrical power is the *on-chip* power consumed by the modulator drivers in transmitters (117μW per driver), and TIAs in receivers (257μW per TIA). For calculating the modulator driver and TIA power we used ITRS device projections [35] and standard circuit procedures. We assume a modulator capacitance of 50fF, even though it is expected to get smaller with technology improvements, and 100fF detector capacitance [55], which is achievable even with current technologies. We also assume a TIA supply voltage that is 20% higher than the nominal supply for our power calculations in the next section.

Optical power is the *off-chip* power required by the modulator to modulate and transmit the information optically from one node to the others. In our analysis, we first calculate the minimum optical threshold power required for a detector to detect a signal correctly, which is based on the voltage swing requirement and signal-to-noise ratio of the receiver as suggested by Connor et al. [55]. In our case, the minimum detector current requirement comes to 30μA. Because only one node transmits with a specific wavelength, and the relative distance between a transmitter and a receiver is known at design time, it is possible to design the detectors to tap only the minimum amount of power adequate for signal detection, resulting in minimum overall optical power. Beginning with the minimum power required at the farthest receiver in the optical loop, we calculate the input power required at the transmitter's modulator by visiting nodes in reverse order up to the transmitter, and accumulating at each step the power losses incurred (Table 3.3). Each modulator requires this amount of optical power, since we assume a continuous wave laser source which will be always on, irrespective of whether data is being transmitted.

We formulate the minimum power per modulator in Equation 3.1. In the

Table 3.3: Major power losses in an on-chip optical transmission system.

	Losses
On-chip coupling loss (dB) [55]	3
Si waveguide loss (dB/cm) [55]	1.3
Splitter loss (dB) [55]	0.2
Modulator insertion loss (dB) [3]	1
Interlayer coupling loss (dB)	1
Bending loss (dB) [55]	0.5
Quantum efficiency [55]	0.8

equation, P_{th} is the minimum power that is required for a detector to detect the optical signal, P_{loss} is the waveguide loss per unit length, L is the length of the bus, and N is the number of nodes on the bus. The first term in the equation accounts for the power required for all detectors, the last term accounts for the waveguide loss, and K accounts for the other losses in the path, such as bending losses, etc.

$$P_{\text{modulator}} = (N-1)P_{th}K \cdot 10^{\frac{P_{loss}L(N-1)}{10N}} \tag{3.1}$$

Using these analytical models, and accounting for the remaining losses in the optical system such as on-chip coupling, splitters, etc., we report the minimum required total optical power for each configuration (Table 3.2). Note, however, that only half of this optical power contributes to the total on-chip power consumption (Table 3.2), as the other half is lost during the coupling of light into the chip (3dB coupling loss).

Discussion

We observe that the most preferable topologies in terms of area and power are H-4x1A{1,2,3}D and H-8x1A1D(4S), although we empirically observe that H-4x1A1D has too low data bandwidth (Section 4.2). All other configurations have excessive power and area expenses in comparison, due to a variety of factors: higher snoop bandwidth, greater number of receivers and transceivers, larger switch crossbars and arbitration logic, etc. Another observation is that, in the four-node configuration, the power consumption of the optical components is relatively low compared to the electrical subnetwork.

Among the preferred organizations, we opt for H-4x1A{2,3}D for our evaluation, mainly because (1) they require lower laser power, and (2) they are more flexible, since they can dynamically allocate the wavelengths for requests from every four L2 caches, while in the eight-node configuration the wavelengths are highly partitioned among nodes, leaving little room for flexibility.

CHAPTER 4

OPTO-ELECTRICAL BUS EVALUATION*

We now provide a first look at the performance impact of incorporating on-chip optical technology for bus-based CMPs. We first present the experimental setup, including the electrical baseline that we model; then, we describe the simulated applications, followed by our results.

4.1 Experimental Setup

We conduct our evaluation using a cycle-accurate execution-driven simulator based on SESC [56]. Latencies and occupancies of all structures are modeled in detail. The simulator models a 64-core chip multiprocessor featuring dynamic superscalar cores and a snoopy-coherent memory subsystem. Each core is 4-way out-of-order and runs at 4GHz. We summarize the core parameters in Table 4.1. Each core has access to a private, write-through L1 data cache. An eight-way banked, write-back L2 cache is shared every four cores through a crossbar. All sixteen L2 caches are connected through a snoopy, fully pipelined bus (the object of our study). The coherence protocol is MESI-based and permits cache-to-cache clean block transfers. A banked, shared L3 resides off chip, but with tags on chip. L3 is accessed in parallel with main memory, and it is exclusive of L2 caches. We model four on-chip L3/memory controllers, each connecting to one fourth of L3 and memory via 64GB/s and 32GB/s links, respectively.

Following common practice for SPLASH-2 applications (Section 4.1.3), we

Table 4.1: Summary of the processor core in the modeled CMP system. In the table, GHR, BTB, and RAS stand for global history register, branch target buffer, and return address stack, respectively. Cycle counts are in processor cycles.

Processor Core	
Frequency	4GHz
Fetch/issue/commit width	4/4/6
Inst. window [(Int+Mem)/FP]	56/48
ROB entries	128
Int/FP registers	96/96
Int ALUs/Branch units	4/2
Int Mul/Div units	1
FP ALUs	3
FP Mul/Div units	2
Ld/St units	2
Ld/St queue entries	24/24
Branch penalty (cycles)	7 (min.)
Store forward delay (cycles)	2
Branch predictor, (Hybrid of GAg + Bimodal)	16K-entry, 14b GHR
BTB size / RAS entries	2048 / 32

use reduced cache sizes to compensate for the applications' reduced working sets [76] as follows: 64×8KB L1, 16×256KB L2, 1×16MB L3. As a sanity check, the last column of Table 4.4 list the global L2 miss rates, as obtained during a bandwidth characterization experiment, which we describe later. Table 4.2 summarizes the memory subsystem parameters.

4.1.1 Electrical Bus

To conduct a meaningful evaluation of the impact of incorporating optical technology to bus-based interconnects, we establish a competitive, state-of-the-art electrical baseline with similar power and active/metal area characteristics as

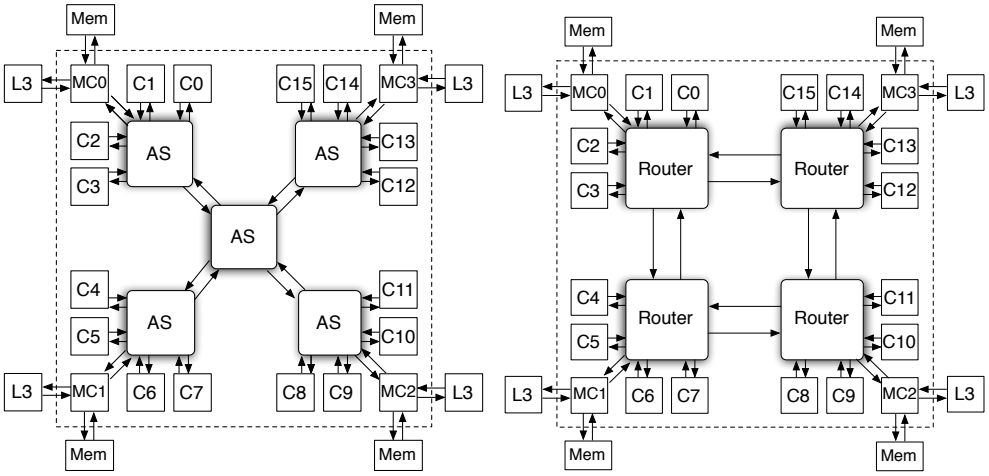
Table 4.2: Summary of the memory subsystem in the modeled CMP system. In the table, MSHR and RT stand for miss status holding register and minimum round-trip time, respectively. Cycle counts are in processor cycles.

Memory Subsystem	
Cache sizes for SPLASH-2 [76]	64×8KB L1, 16×256KB L2, 1×16MB L3
Cache associativity	4-way L1, 8-way L2, 16-way L3
Cache access latencies	2 IL1/DL1, 8 L2, 56 L3 cycles
Writeback/Replacement policy	WT DL1, WB L2 and L3
Block size	64 bytes
MSHR entries	8 IL1/DL1, 32 L2, 12 L3 (per bank)
IL1/DL1 Cache ports	1/3
L2/L3 Cache banks	8/8
L2 Cache coherence protocol	MESI-based
System bus	64 bits, 2GHz
Memory controllers	4
L3 off-chip bandwidth	4×64GB/s
Memory bus bandwidth	4×32GB/s
Memory RT from controllers	320 cycles

the competing opto-electrical buses. We discuss the address network first, followed by the data network.

An address bus can be implemented in a variety of ways, including a hierarchical tree organization (e.g., a single snooping coherence domain in the Sun Fireplane System Interconnect [17] implemented as two-level tree structure), and unidirectional [7, 37, 67] and bidirectional [11] ring-based interconnects.

We empirically found the tree topology to yield low latency and competitive bandwidth relative to other alternatives for our configuration, and therefore choose it as our baseline. We model it after existing proposals [17, 27]. In the modeled tree organization (Figure 4.1a) four L2 caches and a memory controller (which in turn manages one-fourth of the off-chip L3 and memory) connect to an address switch (AS), and four such address switches connect to a top-level



a) Address/Snoop Network b) Data Network

Figure 4.1: Modeled electrical baseline address and data networks. AS, MC(0-3), and C(0-15) stand for address switch, memory controller, and (L2) cache, respectively. Figures are not to scale.

address switch, all through point-to-point links. Requests issued by L2 caches are arbitrated in the switches at each level of the tree, until they reach the top level and are selected. From that point on, broadcasting a snoop request down to all caches, combining snoop responses up at the top-level switch, and again broadcasting the final snoop result down to the caches, takes a fixed amount of cycles. We implement a multibus by selecting multiple snoop requests at the top-level address switch and employing as many snoop request/response buses as needed.

We assume an H-tree layout with 4.5mm first-level (from the L2 caches) and 9mm second-level wire links. By using power-delay optimized repeatered-wires, we can accommodate a 2GHz bus clock frequency–half the cores' speed. Under no contention, the address phase of a request spends a total of 13 bus cycles on the bus: 4 cycles for request arbitration, 3 bus cycles for snoop re-

Table 4.3: Area and power characterization of two possible topologies for the baseline electrical bus, with two and four snoop requests per bus cycle, respectively. Total on-chip power is the sum of all electrical power components. Dynamic power components in switching and wiring columns assume α=1. For α=0.5, only the total sum is provided.

Electrical Topology						
Snoop Requests /Bus clk	**Area (mm^2)**		**Power (W)**			
	Switches/ Routers	Wiring	Switches/ Routers	Wiring	Total On-chip (α=1)	Total On-chip (α=0.5)
2	1.47	15.9	1.42	13.40	14.82	8.08
4	1.66	22.81	1.68	19.23	20.91	11.29

quest, and 6 bus cycles for snoop-response combining and result broadcasting (excluding time spent in the caches).

The data network (Figure 4.1b) consists of a four-node bidirectional ring. As in the case of the address switches, each data router serves requests from/to four local caches and a memory controller connected to it through point-to-point links. Routing is deterministic and balanced. Transfers within a node use a 16GB/s bypass path within the local router. Bandwidth at each ring link is 16GB/s in each direction, as is the read and write bandwidth of each L2 cache. Bandwidth from (to) the memory controller is 48GB/s (32GB/s). In the absence of contention, it takes 14 bus cycles to transfer a cache line on the data network to a cache in the farthest node.

Finally, we do not simulate I/O, and therefore we do not include it in the system we model.

To obtain area and power characteristics of the electrical bus (Table 4.3), we follow the estimation methodology described in Section 3.3.2 for the relevant

electrical components. When compared to H-4x1A{1,2,3}D buses, an electrical bus with support for an equal number of snoop requests per bus cycle (four) exhibits comparable power consumption and active device area, but a 50% increase in metal area overhead. On the other hand, an electrical baseline with support for half as many snoop requests per bus cycle adds up to similar area and power characteristics as the opto-electrical counterparts. Thus, for our comparison, we choose the latter configuration as our baseline.

4.1.2 Opto-electrical Bus

We model the opto-electrical buses H-4x1A{1,2,3}D as described in Section 3.3. The uncontended latencies in these optical buses are 10 bus cycles for arbitration plus snoop request/response phases, and 12 bus cycles for a cache line data to be transferred on the bus across bus nodes.

4.1.3 Applications

We use eleven applications from the SPLASH-2 suite [76] (our simulator currently does not support *volrend*). Their description, as well as their input parameters, are given in Table 4.4. We use MIPS binaries compiled with -O3 optimization level. We fast-forward the initialization part of the applications (at which point we start modeling timing and collecting statistics) and run them to completion.

Table 4.4: Application descriptions and simulated problem sizes. Observed global L2 miss rates (averaged over all L1 and L2 caches) using optimistic (single-bus-cycle address and eight-bus-cycle data transmissions, and no contention) bus (but not memory) are provided for reference.

SPLASH-2	Description	Problem size	L2 miss % 256KB
Barnes	Evolution of galaxies	16k particles	0.15
Cholesky	Cholesky factorization kernel	tk29.O	0.35
FFT	FFT kernel	64k points	1.68
FMM	N-body problem	16K particles	0.10
LU	LU kernel	512×512 matrix, 16×16 blocks	0.12
Ocean	Ocean movements	258×258 ocean	2.35
Radiosity	Iterative diffuse radiosity method	-room -ae 5000.0 -en 0.05 -bf 0.1	0.31
Radix	Integer radix sort kernel	radix 1024, 1M integers	3.47
Raytrace	3-D ray tracing	car	0.80
Water-NSq	Forces and potentials of	512 molecules	0.35
Water-Sp	water molecules (both)	512 molecules	0.07

4.1.4 Bandwidth Characterization

Figure 4.2 plots histograms of the average number of bus requests per processor cycle, sampled at 1,000-cycle intervals, and assuming infinite bus (but not memory) bandwidth, and one- and eight-bus-cycle address and data buses, respectively.

The results show that, for the studied applications, the address/snoop bus bandwidth requirements generally stay below 1.5 req./processor cycle.

4.2 Performance Evaluation

Figure 4.3 shows performance results for H-4x1A1D, H-4x1A2D, and H-4x1A3D, relative to the electrical baseline. Interestingly, in spite of the higher

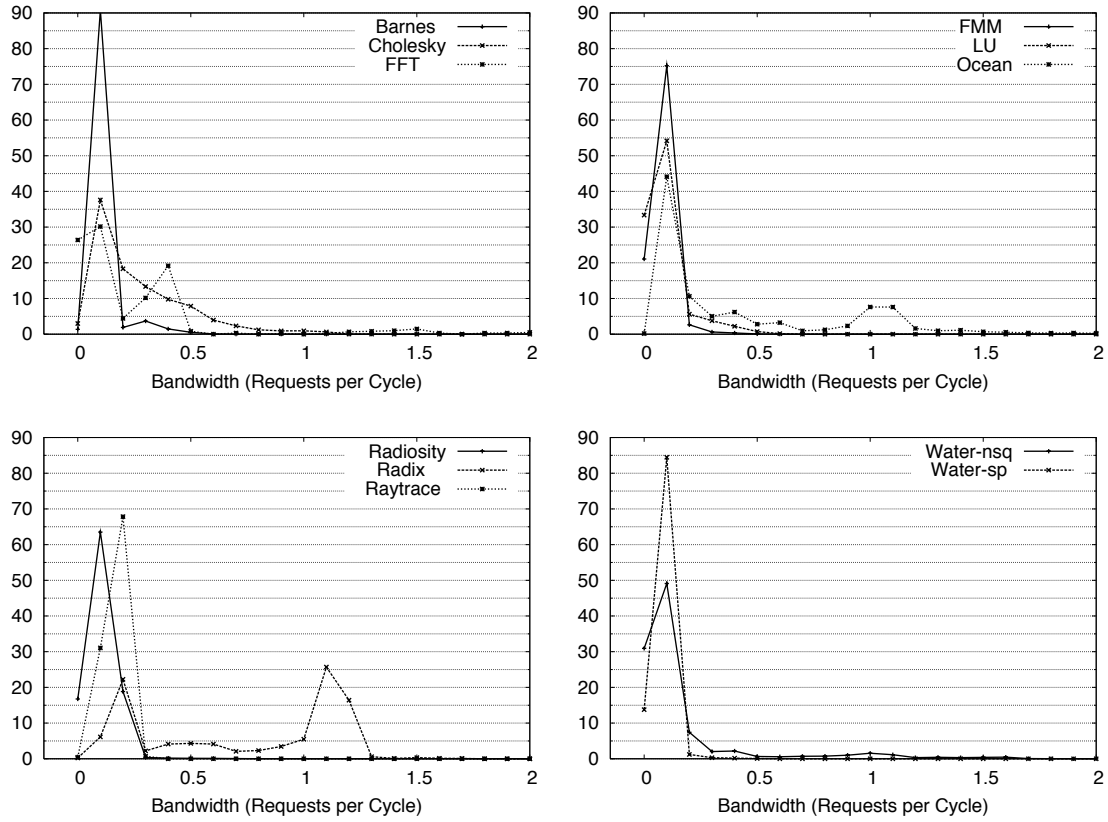


Figure 4.2: Histograms of the average number of bus requests per processor cycle sampled at 1,000-processor cycle execution intervals. An optimistic bus (single-bus-cycle address and eight-bus-cycle data transmissions, no contention) is used.

snoop request bandwidth, H-4x1A1D experiences a significant performance degradation in nearly all cases. This is mainly due to its lower per-node data bandwidth (one outgoing port to the optical bus vs. two outgoing ring-ports in electrical baseline). When higher data bandwidth is provided via additional wavelengths (H-4x1A2D and H-4x1A3D), the opto-electrical configurations achieve significant speedups. The opto-electrical buses can accommodate the L2 miss rates better, resulting in significant speedups: geometric mean of 1.13, an a peak of 1.66 for the H-4x1A3D configuration.

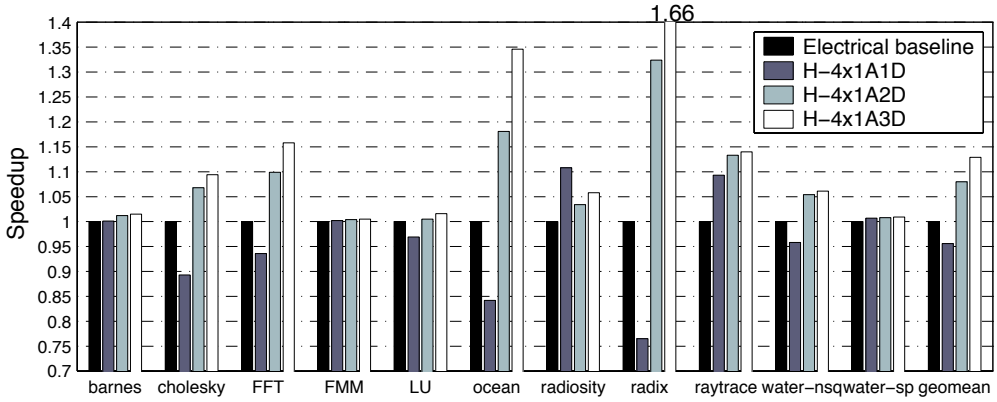


Figure 4.3: Performance improvements of four-node opto-electrical buses as the number of available wavelengths per node for the data network is varied from one to three. (The address network uses one distinct wavelength per node in all three cases.) Speedups are relative to the baseline electrical interconnect.

To further understand the sources of performance improvement, Figure 4.4 shows the average latency breakdown (in bus cycles) of bus transactions in the baseline electrical, H-4x1A2D, and H-4x1A3D configurations. (In the plots, the Data Transfer category excludes memory or cache access times.)

We observe latency advantages for the opto-electrical configurations in both address/snoop and data networks. In the former, effective latency is reduced by 22% on average (34 to 28 bus cycles) when moving from electrical to electro-optical technology. Recall that, even in the absence of contention, the opto-electrical buses have a latency advantage over our electrical baseline. Moreover, the opto-electrical buses can support twice as much snoop request/response bandwidth as the electrical baseline at similar power and area cost. Some applications (*barnes*, *radiosity*, *raytrace*, and *water-spatial*) have significant contention in the arbitration phase. Our simulation data show that this is caused mostly by the serialization of conflicting requests to the same cache line (in our bus

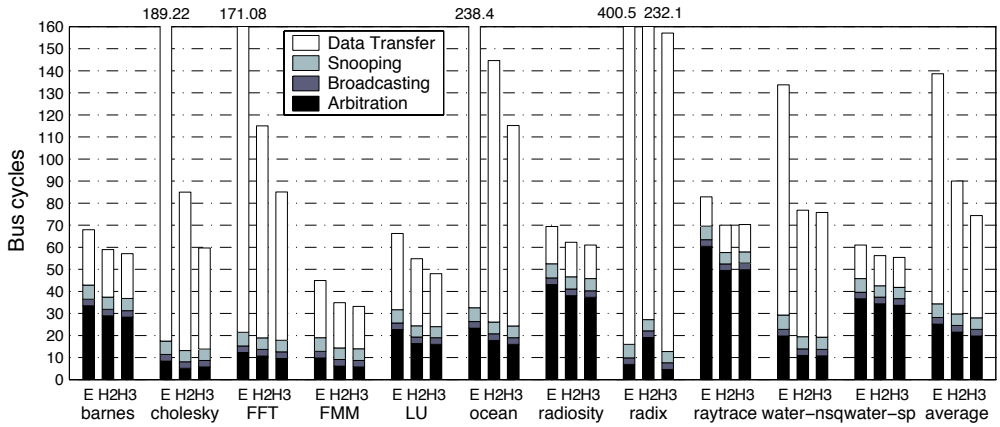


Figure 4.4: Average latency breakdown (in bus cycles) of bus transactions in baseline electrical (E), H-4x1A2D (H2), and H-4x1A3D (H3) buses. Data transfer excludes cache or memory access times.

protocol, conflicting requests to a cache line with an outstanding request are deferred). This is amplified indirectly by the extended latency of the outstanding requests in the data network.

Indeed, for the configurations under study, the main overall benefit comes from reduced contention (and thus effective latency) for data transfers. Our simulations show that the data network struggles to supply the bandwidth needed to satisfy these requests. It is in the data network that the availability of extra wavelengths through WDM yields the largest performance improvements. Still, some applications suffer from significant contention in the data network even for H-4x1A3D, leaving room for further improvement. From our simulation data, we identify the main cause to be contention at the L2 cache input and output ports. Notice that the bandwidth to/from the caches (and memory controller) is kept unchanged in all configurations in spite of the increased data bandwidth on the optical loop. Also, those higher-contention applications would benefit from additional wavelengths.

Table 4.5: Parallel efficiencies of the simulated SPLASH-2 applications for the specified configurations.

SPLASH-2	Baseline	H-4x1A2D	H-4x1A3D
Barnes	0.84	0.85	0.85
Cholesky	0.34	0.36	0.37
FFT	0.46	0.50	0.53
FMM	0.61	0.62	0.62
LU	0.51	0.51	0.51
Ocean	0.48	0.55	0.63
Radiosity	0.30	0.31	0.31
Radix	0.28	0.37	0.46
Raytrace	0.18	0.21	0.21
Water-NSq	0.72	0.76	0.77
Water-Sp	0.84	0.84	0.84

Finally, Table 4.5 shows parallel efficiencies (relative to a sequential run in the same configuration in each case) for all applications running on the electrical baseline and on H-4x1A{2,3}D. In general, scalability improves with the addition of optical technology. Not surprisingly, those applications that suffer from more contention in the data network tend to exhibit lower parallel efficiencies in all configurations. And it is precisely the scalability of these applications that improves the most with the addition of optical technology.

In summary, our evaluation shows that incorporating optical technology in bus-based CMPs can have a beneficial impact on performance, and that WDM support may be critical to effect this impact in both address/snoop and data networks. The fact that WDM comes at very small additional area and power is encouraging. In the particular design points that we evaluated, the contribution to performance by the data network turned out to be dominant.

CHAPTER 5

ALL-OPTICAL NETWORK USING WAVELENGTH-BASED OBLIVIOUS ROUTING*

Our preliminary work employs broadcast-based data communication on a full optical crossbar. It is a high-bandwidth, low-latency organization which does not require global arbitration. However, the $O(N^2)$ detector/receiver requirement is likely to be an issue for high node counts (N), in terms of sheer component count and the complexity involved in processing all the messages a node can receive simultaneously. A fully-optical implementation of the design is inviable due to excessive power requirements. In the final solution, we rein in this problem by resorting to a clustered electro-optical organization that reduces the number of nodes at the optical crossbar. The downside of a clustered electro-optical approach is that its potential may be limited by the latency and power requirements of the electrical side.

We strongly believe that an all-optical approach to constructing an on-chip data network constitutes a very attractive proposition from the performance standpoint, and that a careful design can deliver a fundamentally power-efficient solution that is reasonably robust to technology considerations. In the second part of the dissertation, we argue for such an approach. Specifically, our proposed optical-routing based solution combines the following key features:

Wavelength-based routing. Within each optical router, the route followed by a packet depends solely on the wavelength of its carrier signal, and not on information either contained in the packet or traveling along with it. This allows us

to adopt an all-optical solution for data transmission, where O-E/E-O support at each router to figure out routes is unnecessary. Wavelength-based routing is a popular approach in optical LAN/WAN technology for this same reason [87].

Oblivious routing. The wavelength (and thus the route) employed to connect a source-destination pair is invariant for that pair, and does not depend on ongoing transmissions by other nodes, thereby simplifying design and operation.

Passive optical routers. Their routing pattern is set at design time, which allows for area and power optimizations not generally available to solutions that use dynamic routing. It also means no time lost in routing/arbitration decisions.

We construct such an all-optical network layer where each node has physical connectivity to all other nodes via static (wavelength, waveguide) paths and a common input and output port. We replicate this network layer to increase bandwidth. We employ connection-based operation to better exploit the network. A source node first establishes a logical connection with the destination node before transmitting a data packet. Then it maintains the connection for as long as possible to transmit data without additional global arbitration overheads. A node may have concurrent connections to multiple nodes both on the same and different network layers. We also propose techniques to hide connection establishment delays and to increase connection utilization.

Connection-based operation of the data network can benefit applications by forming logical connections on the network layers that match applications' communication pattern, thus minimizing global arbitration and streamlining data transfers. Our data network design also provides good isolation between exclusively communicating groups of nodes.

First, we construct an oblivious, wavelength-routed, all-optical network for CMPs using nanophotonic components and describe its connection-based operation. Then we evaluate the performance of the proposed network in the context of a shared-memory 64-core, 256-thread CMP design in Chapter 6. Finally, we analyze the design in terms of cost and power in Chapter 7.

5.1 CMP Architecture

The CMP architecture of our study comprises 64 2-issue, in-order, 4-way multithreaded cores with their private L1 i- and d-caches. Each core is augmented with 4-way SIMD support, providing 16 GFLOP/s peak performance at 4 GHz core frequency, for an aggregate peak CMP performance of 1 TFLOP/s. Cores are organized in clusters of four, and cores within each cluster share a L2 cache. The system further employs eight memory controllers, each providing access to one of eight cache-block-interleaved L3 cache + memory banks. Each controller can deliver up to 256 GB/s.[1]

The shared-memory system maintains coherence across L2 caches and lower level L3 cache and memory, using a MESI-based snoopy protocol, and a pipelined split-transaction opto-electrical command/snoop bus along the lines of [39] that runs at processor frequency. Actual transfer of cache blocks takes place in the data network, which is the subject of our study. In the following sections, we describe the design and operation of an oblivious, wavelength-routed, all-optical data network that interconnects the sixteen L2-cache nodes and the eight memory-controller nodes. Section 6.1 provides more details on the CMP architecture.

[1]Preliminary estimations indicated, and later simulations confirmed, that this provisioning is adequate to support the bandwidth demand of the 64 cores.

5.2 Network Substrate

In wavelength-based routing, the route a packet takes at each point in the network depends exclusively on the wavelength of its carrier signal. This is advantageous because it allows us to offer end-to-end optical data transmission, without having to undertake OE-EO conversions and buffering in order to route a packet based on its content.

Oblivious routing, on the other hand, dictates that a given source-destination pair always communicates via a predetermined wavelength, which does not depend on the ongoing transmissions between other source-destination pairs. It enables us to provide connectivity using passive optical routers on the network, based on preset microring resonators that will automatically route each wavelength on the right path to the destination.

Ideally, one could conduct a multi-dimensional design space exploration (topology, routing, etc.) to devise a network that simultaneously optimizes for cost, complexity, and performance. For simplicity, however, in this work we pick a reasonable waveguide and optical router topology, and then work out a viable routing scheme. Specifically, after some preliminary trials, we opt for a 24-node two-dimensional torus. A two-dimensional torus is attractive because, as we will see later, it yields relatively simple routers and waveguide layout. Also, as we explain next, it is a viable topology for wavelength-based oblivious routing.

With these in mind, we set out to devise a wavelength assignment for all source-destination pairs, followed by a search of (passive) router configurations, that will result in a viable routing solution with full connectivity.

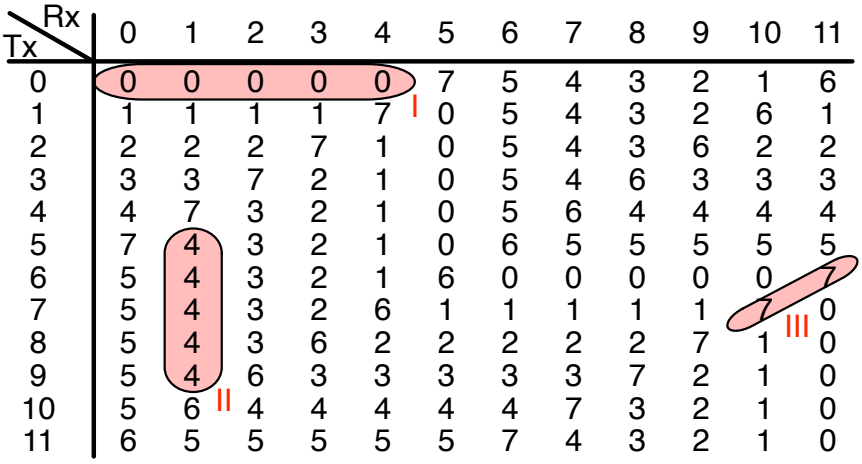


Figure 5.1: Optimal wavelength assignment found using Aggarwal et al. [1] for oblivious routing in 12-node wavelength-routed optical network. (i, j). element in the matrix gives the wavelength that must be used when node i needs to communicate to node j. 8 wavelengths (labeled 0 through 7) are required. Highlighted cases I, II, and III show examples of wavelength reuse.

5.2.1 Wavelength assignment

In oblivious routing, every source-destination pair must have an assigned wavelength through which to communicate. A trivial way to accomplish this is to employ as many wavelengths as the number of distinct source-destination pairs. This, however, not only is prohibitively expensive ($O(N^2)$ wavelengths, where N is the number of nodes), but also unnecessary. Indeed, Aggarwal et al. [1] prove that significant wavelength reuse is possible. Specifically, the number of wavelengths needed to support oblivious routing in a network with N nodes is $(\lceil \frac{N}{2} \rceil + 2)$ for $N = 4$ or $N \geq 6$, assuming that, at any time, source nodes communicate to different destination nodes one-to-one.[2] The authors further provide an

[2]The authors also assume that each node is connected to a router through a pair of incoming and outgoing physical channels, in our case waveguides. However, the authors ignore the complexity of the routers and the connectivity between them. As we discuss next, our problem is

algorithm to calculate the wavelength assignment to connections between pairs.

Figure 5.1 shows a solution for a 12-node network using eight wavelengths (labeled 0 through 7). Element (i,j) in the matrix contains the wavelength that must be used when node i needs to communicate to node j. There is notable wavelength reuse: A source node uses the same wavelength to communicate to several nodes (case I); multiple source nodes use the same wavelength to communicate to the same node (case II); distinct source-destination pairs also use the same wavelength (case III). Nevertheless, the wavelength assignment is such that one-to-one communication between distinct source-destination pairs can concurrently take place without conflict at any of the receivers, which cannot discriminate between messages sent over identical wavelength. Figure 5.2 shows a few simple scenarios where careful assignment is in order. In the leftmost one (case I), for example, node A uses wavelength w_i when transmitting to either node X or node Y. It follows that any other node B necessarily uses $w_k \neq w_i$ to communicate to Y. Otherwise, when A transmits to X and B transmits to Y concurrently, A's and B's signal would interfere at Y. The figure provides assignment constraints for case II and III as well.

We use the algorithm by Aggarwal et al. [1] to obtain the wavelength assignment for our 24-node system.

5.2.2 Wavelength-path layout

Once we derive a wavelength assignment for all source-destination pairs, we must determine the exact wavelength paths on the torus network, which comes

more restrictive than this, since our network is physically constrained and communication pairs do share physical medium often.

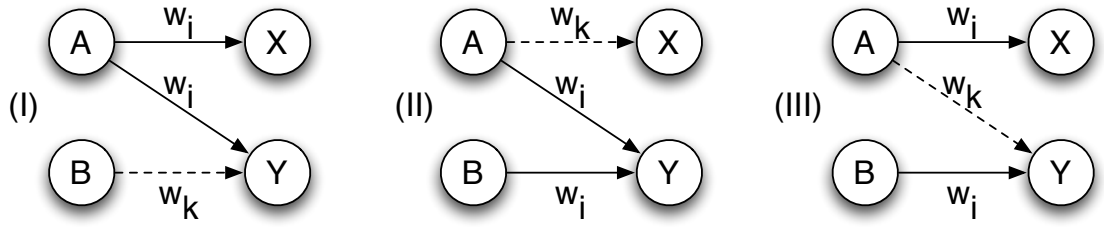


Figure 5.2: Three examples of wavelength assignments where w_k must be different from w_i to guarantee interference-free reception. For simplicity, the $B \rightarrow X$ wavelength assignment is not shown.

to determining the static routing configuration of the wavelength routers. A wavelength sourced from a node should only reach the destinations designated by the assignment, and not the others. The limited number of waveguide segments on the network makes it challenging to satisfy this routing constraint. In the worst case, it may not be possible to map the wavelength assignment. We must search the configuration space and find one which successfully routes the wavelengths from all nodes, necessarily without wavelength collisions in waveguides. Notice that, because we are using fully optical transmission, non-minimal routes are not necessarily a concern, and in fact they are attractive to the extent that they may enable a successful routing.

A manual search would be prohibitively time consuming and error prone. For this reason, we implement a genetic algorithm (GA) to automatically find a viable configuration. We solve the problem one wavelength at a time observing that solutions for different wavelengths are independent. Our GA begins with a set of randomly-generated configurations. It works its way toward a solution by applying a multi-objective fitness function that rewards routes that are closer from successfully connecting a source-destination pair and have hop count closer to the minimum possible hop count, and penalizes undesirable out-

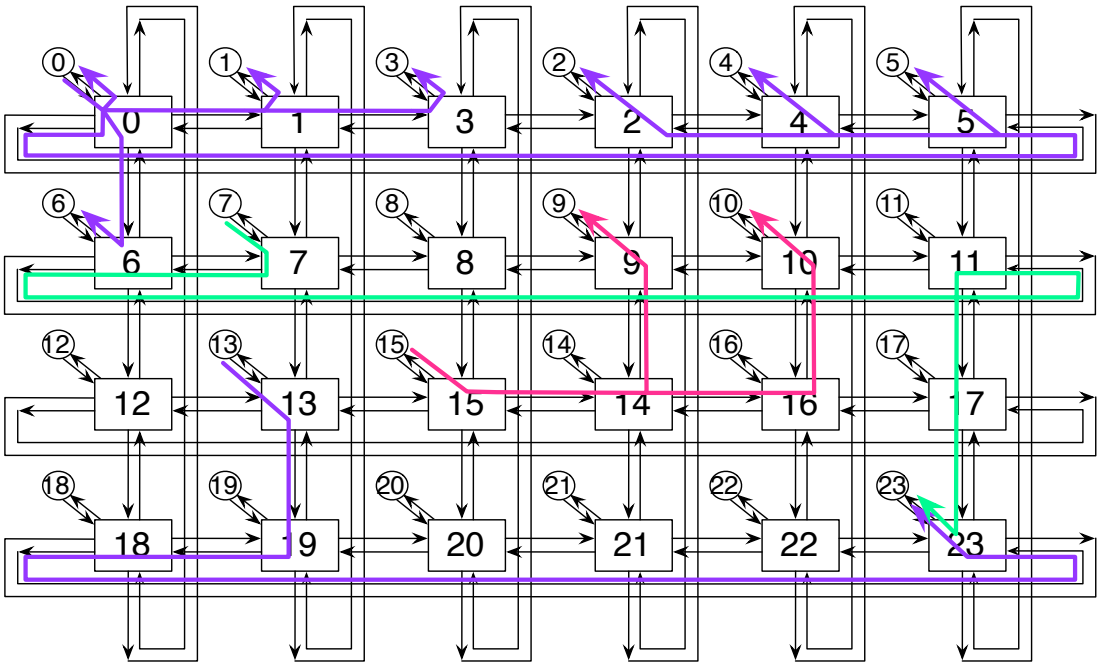


Figure 5.3: The 6x4 two-dimensional torus adopted in this study. Notice the shuffling of a few node labels with respect to "typical" labelings, which was needed for the GA algorithm to find a viable solution. Several routing paths from the actual solution are shown.

comes, such as routes with loops and routes that reach unwanted destinations as a result of collisions with other routes. As soon as the GA is able to find a viable solution, the search stops.

For "typical" labeling of torus nodes (e.g., XY labeling), our GA was not able to find a viable solution for the two of the wavelengths. We considered adding node labeling to the search space, however we found that a few simple label swaps by hand (nodes 2 and 3, and nodes 14 and 15) were sufficient to steer the GA toward viable solutions. Among complete solutions from multiple GA runs that also satisfy the minimum number of hop counts, we picked the one which resulted in lower optical power and smaller propagation distances. While we

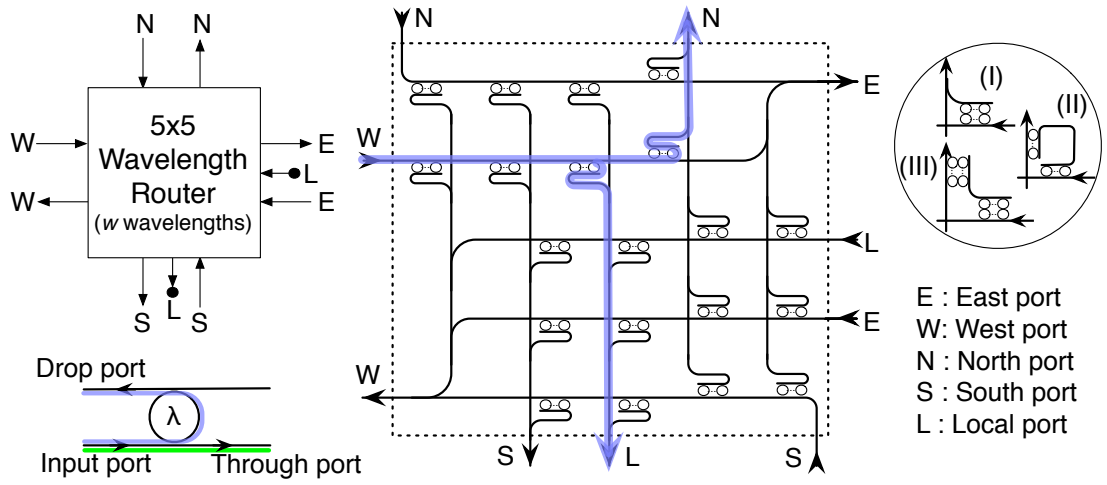


Figure 5.4: Passive wavelength-router implementation. A few alternative junction implementations are also shown (I, II, and III).

could have incorporated those features to the GA's fitness function, we were sufficiently satisfied with the solutions at hand that we did not pursue that for this work. Figure 5.3 shows the labeling of the nodes, as well as a few routes contained in the final configuration. We provide the full listing of all the routes in Appendix B.

Once the routes are determined, each router is customized at design time to satisfy these. We discuss router design next.

5.2.3 Wavelength-router design

We construct passive wavelength routers as depicted in Figure 5.4. Routing a wavelength from an input to an output port is accomplished via careful placement of a passive microring resonant to the wavelength at the appropriate input-output waveguide junction. In a junction, there are as many microrings

as the number of wavelengths that are routed from the input to the output port.

A microring-resonator based filter is reviewed in the lower left corner of Figure 5.4. It is an optical component whose geometry (e.g. radius, separation with the neighbor waveguides) determines the resonance wavelength and coupling ratio of the filter. Light from the input port passes the microring and continues on the through port if the wavelength of the beam and that of the microring do not match. If they match, the light is coupled to the ring and dropped at the output waveguide. Based on the coupling ratio, a fraction of the light may continue on the through port.

Figure 5.4 illustrates a few alternative junction implementations as well. Notice that the routers may be completely different from one another, as needed to implement all the routes found by the GA. The resulting router designs are rather compact, with 2.06 microrings per junction on average (8 maximum). This is encouraging in terms of potential area and power savings.

5.2.4 Transmitter/receiver interface

Recall that, in the wavelength-assignment formulation described, nodes may not be transmitting/receiving to/from more than one node at any point in time. Consequently, we restrict each node to have a single input and a single output port from/to the network. On a transmission, the source must select and modulate on the node's assigned wavelength for the intended destination. Likewise, the destination must select and detect the same wavelength.[3] To accomplish this, we implement wavelength filters at both ends using an array of active mi-

[3]This will require a protocol to have source and destination nodes tune to their assigned wavelength prior to transmitting data. We explain one such protocol in the next section.

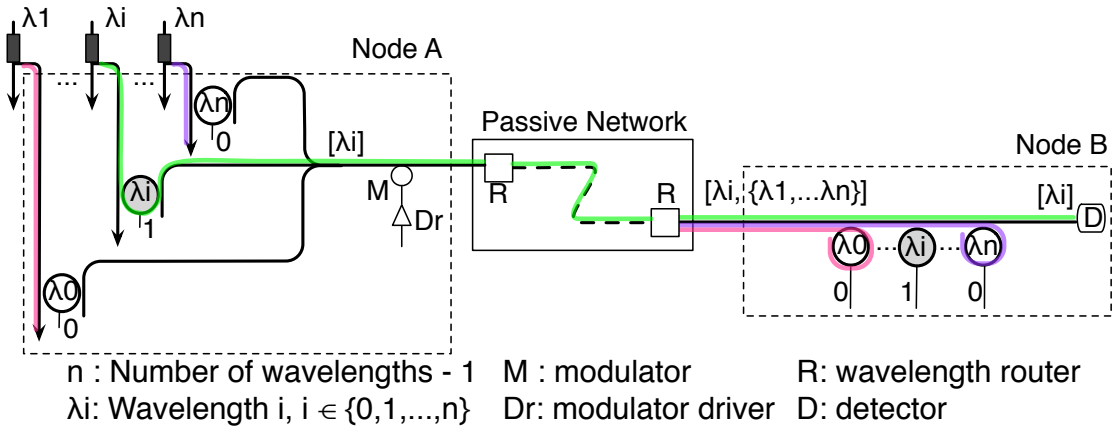


Figure 5.5: Simplified diagram of interfaces for transmitting and receiving data at end nodes. In the figure, node A is transmitting information to node B over wavelength λ_i.

crorings, with a microring per wavelength (Figure 5.5). We assume separate waveguides distribute wavelengths to nodes to optimize for power, because wavelengths have different light paths, hence power requirements. When there is no transmission, the microrings are off-resonance (on-resonance) by default at source (destination) filter. Therefore, wavelengths pass (couple into) the rings and are not injected into (extracted from) the network. On a transmission, only one of the rings is shifted on-resonance (off-resonance) by exerting power, allowing the corresponding wavelength to couple into the input waveguide at the source side or pass to the detector at the destination. Simple decoders can be used to drive the microrings. With this organization, tuning can be very fast. Notice also that tuning need only change when a source/destination node must move to another wavelength, in order to participate in a data exchange with a different node. Finally, we use a modulator and detector that can work with whatever wavelength is offered by the preceding filter.

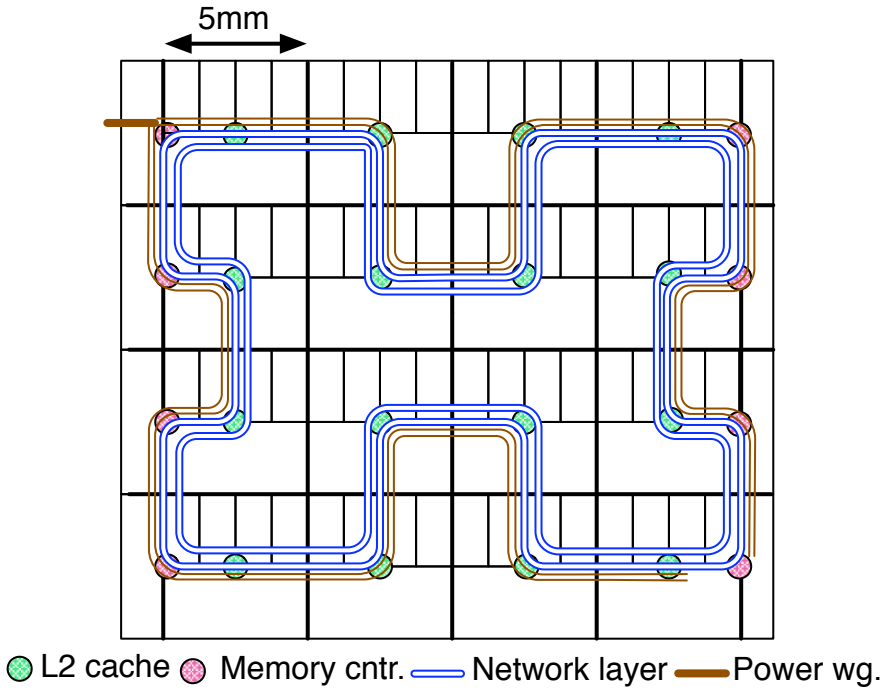


Figure 5.6: Concentric layout of multiple network layers.

5.2.5 Multiple network layers

The single oblivious-, wavelength-routed network layer discussed so far enables one transmission at optical modulation rate from each node at a time. A cost-effective way to augment the network's bandwidth is to embed multiple *virtual* networks in the same set of waveguides, using spare wavelengths which may be available depending on the maturity of the technology. One possibility is to employ the technique proposed in Small et al. [64], which essentially places several wavelengths in the resonance band of a microring resonator. In that case, it is possible to route multiple bits of a message in parallel with little extra hardware: At each node, multiple modulators/detectors must tap separately on each of these wavelengths in order to inject/extract the bits of information; however at the routers and filters the only change comes from broadening the

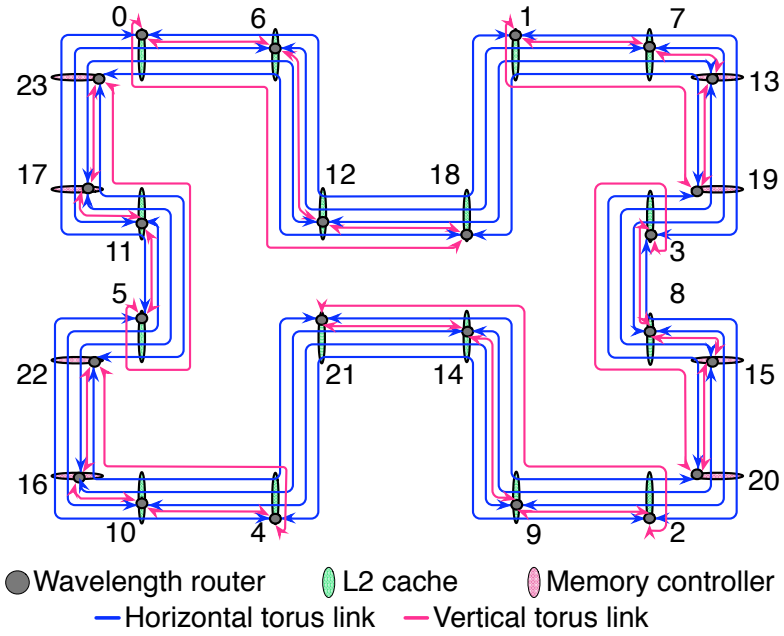


Figure 5.7: Circular layout of a torus network layer. The numbering is matched with the node/router numbering in Figure 5.3. For clarity, we draw the two unidirectional links between two routers as a bidirectional link.

resonance band of their microrings, in order to correctly route such wavelength bundles.

Another way to achieve higher network bandwidth is simply to replicate the network. Notice that all such network layers must be laid out in a manner that minimizes waveguide crossings, which are a major source of optical power losses. In our design, we lay out the network layers in a circular and concentric fashion (Figure 5.6). Each layer still has torus connectivity (Figure 5.7).

Multiple physical network layers can be used to transfer more bits of the same message, or alternatively, more messages. A node splits enough optical power from each wavelength to feed the maximum number of network layers that can transmit concurrently on that wavelength. This input light power is

distributed to network layers based on demand by the active filters. In case multiple physical network layers transfer more bits of the same message, these layers share a source-side filter in a node. After the filter selects the appropriate wavelength based on the target destination, common to all of these layers, the light at the filter output is split to individual layers for modulation.

Unless otherwise stated, in the network operation below, we assume all layers are used to transmit different messages.

5.3 Network Operation

The formulation and physical design of a network layer dictates that a source-destination pair tunes to the assigned wavelength before the actual transmission takes place, and only one source node attempts to transmit to a destination node at a time on a layer. The latter essentially requires arbitration for a receiver.

We devise a simple, distributed protocol that not only ensures these constraints are satisfied for a single data transmission, but also retains this *contract* for as long as possible so that the source can transmit data at any time later without incurring additional arbitration delays. During the *connection* established through such a contract, the receiver necessarily remains tuned to the assigned wavelength, while the transmitter needs only tune right before a data transmission. This allows time multiplexing the transmitter for different connection uses. A connection lasts until it is closed by the destination node, for example when it needs to participate in a connection with another node. A node can own multiple connections to different nodes both on the same layer and on different layers. In this work, we impose the limitation that a source-destination pair can

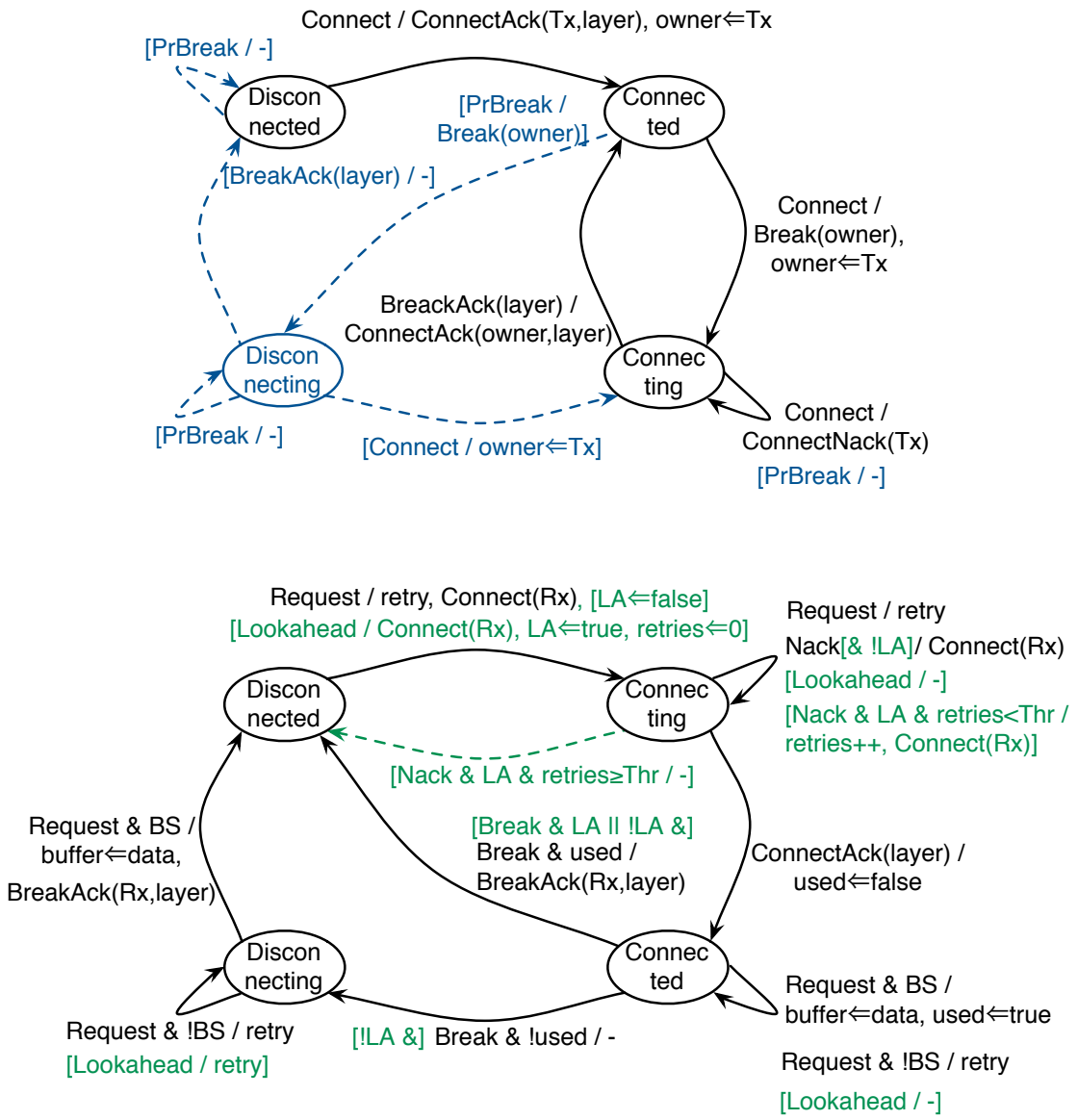


Figure 5.8: Protocol state diagrams for Rx-side (top) and Tx-side (bottom). BS and LA stand for buffer space and lookahead flag, respectively. Protocol extensions for proactive break and connection-lookahead support are encapsulated in [ ] brackets and use dashed transition lines. The information in ( ) specifies the destination of the message and/or extra information received/delivered with the message, such as layer information.

be connected only on one (but any) of the network layers at a time.

The connection protocol is carried on a few dedicated optical network layers via point-to-point transactions between nodes. Clearly, its operation is not connection based. We use a time-slotting approach to ensure that the above constraints are satisfied for a single message transmission at a time. Fortunately, the small size of the protocol messages and the frequency of these transactions require less bandwidth, rendering this approach a viable one.

In the following sections, we describe the connection protocol, network-layer selection policy, the operation of the protocol network layers, and finally the hardware support at the network interfaces.

5.3.1 Connection Protocol

For simplicity, we first describe the connection protocol assuming a single network layer between nodes.

A source node issues a connection request (*Connect*) to a destination node if it finds it is disconnected to that node on a data transfer attempt. In the simplest case, the destination directly acknowledges the *Connect* if its receiver is disconnected as well. If the receiver is involved in a connection, on the other hand, the destination node first needs to break that connection, and wait for any scheduled transmissions by the previous owner to complete before sending the connection acknowledgement to the new connection requester. Once the *Connect* requester receives the acknowledgement, it can start sending data at any time, without any consideration of other nodes. The previous owner, on

the other hand, would need to establish a new connection before any future data transmission to the destination node.

The full protocol is slightly more involved, due to issues that arise from the nonatomicity of the connection setup process. Below we briefly discuss these issues.

Connection-request races: Multiple connection requests can compete for a receiver in a node. The node is the point of synchronization: upon accepting the first request, it will not accept further requests until the outstanding one is resolved from the node's point of view. Requests that find the receiver busy connecting are nacked and retried later.[4]

Forward progress: In order to avoid receiver ownership to ping-pong between nodes without actually being used, a receiver owner delays the break acknowledgement until the connection is used at least once. Because the connection is established in response to a data transmission attempt, it is guaranteed that the connection will be used at least once.

Scheduled transmissions: At the time a connection owner receives a break request, it may have scheduled data packets on this connection in the transmitter's buffer. The break acknowledgement is piggybacked to the last of these packets. The connection is closed from the point of view of the owner. In case of no scheduled transmissions, and the connection has been used at least once, the break is directly acknowledged on the protocol network layers. If the connection has not been used, the break acknowledgement will be piggybacked to the first data packet to be scheduled on this connection. As a result, a break ac-

[4]There are different techniques to avoid starvation for nacked requests [25].

knowledgement for a connection always reaches the destination after all packets scheduled on this connection.

Reply-request races: A connection acknowledgement and a subsequent break request for the same connection, could potentially overlap in the network. The protocol-network-layer interface guarantees that the reply is delivered before the request (Section 5.3.3). This simplifies the protocol.

Similarly, a break acknowledgement and a subsequent *Connect* request for the same source-destination pair can overlap in time. Although ordered delivery of the reply and the subsequent request to the same node is guaranteed on the protocol network layers, recall that, a break acknowledgement can be delivered over a data network layer, possibly after the *Connect* request. This does not constitute a problem, because the *Connect* request will find the receiver busy connecting and will be nacked.

Figure 5.8 summarizes all protocol actions in two state diagrams–one for the receiver side and one for the transmitter side. The diagram also shows protocol extensions to support a few performance optimizations that we describe later.

5.3.2 Network-Layer Selection

The main challenge with multiple network layers is to decide on which layer to establish a new connection.

On a *Connect* request to a node, the node applies a selection policy to choose a network layer on which to establish the connection. This may result in evicting an existing connection. This is conceptually similar to victim selection in a cache

replacement policy. The selection policy that we implement in this work is LRU; we tried others (round-robin, random, etc.) and found their performance to be at most as good as LRU's. Once a layer is selected, the connection protocol is executed for this layer. The layer information needs to be communicated in the relevant protocol transactions, which we already include in the state diagram in Figure 5.8.

A data transmission necessarily takes place on the layer with an established connection to the destination node. A node keeps track of connection status to each destination separately. The status information include the layer id.

Notice that, in the case of unordered delivery of break acknowledgement and a subsequent *Connect* request for the same source-destination pair, a different layer may be selected for the new connection while the previous connection is currently being disconnected on another layer. Nevertheless, the previous connection had been closed from the point of view of the source node before sending the new *Connect*.

5.3.3 Protocol Network Layers

A node transmits the protocol messages for any of its connections, either as a source or destination, on a few dedicated network layers described in Section 5.2. A deterministic and periodic time-slot schedule dictates to which nodes it can transmit every time slot on these network layers. Each layer is always used to transmit to the same set of nodes. Thus, a node can send message to a particular node every $\frac{N}{M}$ time slots, where N is the number of nodes, and M is the number of protocol network layers. Note that the schedules are shifted

from one node to another to ensure that only one node attempts to transmit to each node at a time slot. A node has a schedule for each of its receivers on the protocol network layers as well, so that they can properly tune each time slot. A time slot should be long enough for a protocol message to reach its destination node. It should also accommodate the tuning delays. Based on the time-slot schedules, some light paths on each protocol layer will not be used. Obsolete components can be removed, potentially reducing component counts and power. We perform this optimization in our power evaluation (Section 7).

A node processes incoming protocol messages in a non-blocking and pipelined fashion.[5] It places outgoing protocol messages in an output buffer per layer with an entry for each possible destination. Messages wait here for their time slots.[6] Because very little information needs to be stored in each entry, the overall storage overhead is small.

5.3.4 Hardware Support

Figure 5.9 depicts a node's interface to the optical network. A connection-status table tracks the outgoing connections to each node, while a receiver-status table holds connection information for the receiver on each network layer.

A data transmission attempt first checks the connection status for the destination. If the connection is open, the data is placed into the transmitter buffer

[5]Multiple cycles in a time slot allow lower port requirements to process the incoming messages in a slot.

[6]Note that, there may be multiple protocol messages waiting for a time slot. However, because of the way the protocol and its network layers work, there can be no two messages of the same type (connection request, connection ack/nack, break request, and break ack) targeted to the same destination node. Therefore, an entry has separate fields for the four protocol-message types. Protocol-message arbitration ensures that reply-request races (Section 5.3.1) for the same destination are properly ordered.

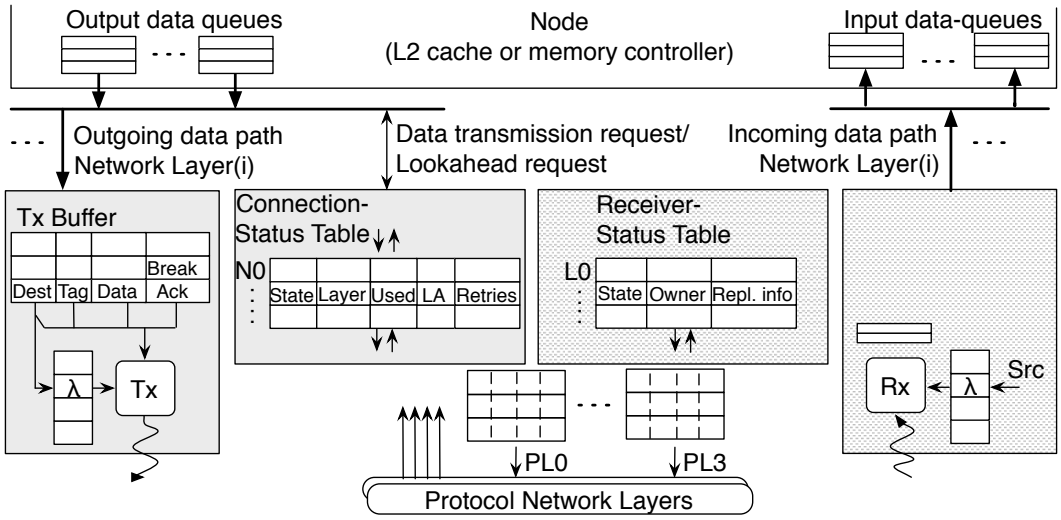


Figure 5.9: Node's interface to data (L) and protocol (PL) network layers. LA stands for lookahead.

of the connection's layer. The connection usage information is updated. In case the transmitter buffer is full, the data transmission attempt is retried later. If the connection is not ready or does not exist, the transmission attempt is delayed until the connection is established.

Newly generated protocol messages are scheduled for transmission on a protocol network layer. They wait a proper time slot in the layer's outgoing buffer (Section 5.3.3).

Protocol messages received from the protocol network layers are processed using either connection- or receiver-status table depending on whether the message is directed to an outgoing or incoming connection.[7]

On an actual data transmission from the FIFO transmitter buffer on a network layer[8], the transmitter first tunes to the destination's wavelength, which

[7]Note that, for correct operation, a break-request processing must see the simultaneous connection use from the node.

[8]If a set of network layers is used to transmit bits of the same message in parallel, these layers

it looks up from a wavelength mapping table, and then transmits the data. Notice that, the node may have multiple connections on this layer that use the same wavelength (recall case I in Table 5.1). In this case, a data transmission will reach all receivers for these connections. A receiver, therefore, checks the intended destination of the data packet before delivering it to the node. We employ non-blocking delivery to the node accomplished through matching the total delivery bandwidth to the aggregate receive bandwidth from the network layers, and employing double buffering where one of the entry is filled until the other entry is delivered. The command/snoop phase preceding the data transmission guarantees that there is available input-buffer entry at the node. We also assume critical-word-first delivery. For systems where buffer space at destination is not guaranteed, or delivery rate does not match the receive rate, credit-based control-flow support can be easily added by leveraging the protocol network layers to communicate the credits. We discuss how to achieve this in Appendix C.

5.3.5 Optimizations

Here we discuss a number of possible optimizations, and any protocol or hardware changes when required.

Lookahead connection requests

In the basic protocol, a node requests a connection only on an actual data-transmission attempt. It is possible for a node to act earlier for establishing a

will share the same transmitter buffer.

connection in anticipation of a future data transmission. The hope is to hide connection establishment latency. There are several circumstances when we apply this feature:

– A memory controller issues a lookahead request when it sends a read request to the L3 cache and memory, so it can relay the data promptly to the requesting node once it returns.

– A cache with E or M state in a snoopy MESI-based coherence protocol issues a lookahead request concurrently to sending its snoop response, in preparation for the data transmission that will follow shortly.

– On a cache line eviction, a lookahead request is issued in parallel to sending the write-back request through the command/snoop bus (this assumes that the cache knows the memory bank's location in the network).

Figure 5.8 shows the protocol changes on the transmitter's side needed to support this feature; the protocol in the receiver's side is unchanged. Note that, because it is not guaranteed that a connection established through lookahead is going to be ever used, the protocol directly breaks such a connection on a break request if no transmissions are pending. Also, we drop a lookahead request after being negative acknowledged for a certain number of times.

Connection-aware cache coherence

In the context of the coherence protocol, upon a read/read with write intent request by a node for a cache line in Shared state at one or more remote nodes, a

subset of the sharers may already have an established connection with the data requester node (necessarily on different network layers). We propose that one of these sharers leverages the existing connection and provide the data. Such sharer nodes should include the fact that they have an open connection in their snoop response, so that the coherence controller may consider them as preferred suppliers.

Proactive connection-break requests

In the basic protocol, a node breaks a connection only in response to connection requests by other nodes (Section 5.3.1). We extend the protocol so that a node can proactively initiate the break of one of its incoming connections. The hope is to hide the break-handshake latency on a subsequent connection establishment. If a connection request reaches the node before it has successfully broken the connection, it need not resend the break request; the request simply waits for the acknowledgment from the current receiver owner. Figure 5.8 shows the protocol changes on the receiver's side needed to support this feature; the transmitter's side does not change.

In our implementation, a node triggers proactive break upon processing a connection request. It uses LRU policy to select a used connection to break proactively. Depending on the particular network configuration, some nodes may not trigger proactive break at all, while some nodes may proactively break multiple connections each time. We describe these cases in our experimental setup.

CHAPTER 6

PERFORMANCE EVALUATION OF OPTICAL NETWORKS*

This chapter analyzes the performance of the proposed network, in the context of a 64-core 256-threaded CMP targeted for 32 nm technology node, compared against alternative designs proposed in the literature. Later in Chapter 7, we perform detailed power analyzes for the evaluated networks.

6.1 Experimental Setup

This section provides more details on the evaluated CMP architecture whose overall organization we highlighted in Section 5.1. Tables 6.1 and 6.2 summarizes core and memory-system parameters.[1] We use CACTI 5.3 [71] to obtain cache latencies. We assume a 450 mm^2 die area, which is in line with server-oriented CPUs.

Following common practice for SPLASH-2 applications, we use reduced L2 cache size of 256 KB to compensate for the applications' small working sets [76]. Still, we use the latency of an L2 cache's full-size equivalent of 2 MB.

Banked L3 cache is on a separate 3D layer. 3D interconnection provides 256 GB/s bandwidth from each bank. High off-chip memory bandwidth of 2 TB/s (256 GB/s per memory bank) is provided through optical chip-to-chip interconnection. Optical access to memory arrays reduces the memory latency

[1] In our simulation infrastructure, 4-way SIMD processing is emulated by issuing up to 4 consecutive independent floating-point add/sub/mult instructions with ready operands in one cycle. Any intervening instruction in the code not of one of these types terminates the SIMD bundle.

Table 6.1: Processor core of the modeled system. In the table, GHR, BHR, BTB, MSHR, and RAS stand for global history register, branch history register, branch target buffer, miss status holding register, and return address stack, respectively. Cycle counts are in processor cycles. Bus latencies are contention-less latencies.

Processor Core	
Frequency	4GHz
Issue	2-way in-order
Int ALUs/Branch units	2
Ld/St units	1
Mul/Div units	1
FP ALU/MUL units	4-way SIMD
FP Div units	1
Write-buffer entries	16
Store forward delay	2 cycles
Branch min. cycles	5
Branch predictor,	13b GHR, 2K 10b BHRs
(Hybrid of GAg + SAg)	8K Chooser
BTB/RAS entries	2048/32
IL1/ DL1 size, associativity	32KB, 4-way
IL1/ DL1 access latency	2 cycles
IL1/ DL1 block size	64B
DL1 writeback policy	WT
DL1 MSHR entries	16

as well [72, 8].

We evaluate several configurations of the proposed network. We also compare them against two optical networks modeled after previously proposed architectures [39, 72].

Bisection bandwidth in all configurations is set to 6 TB/s as an equalizing parameter to make meaningful power and bandwidth-utilization comparisons. For all configurations, we assume support for up to 64 wavelengths [72], 10.45 ps/mm light propagation delay [39], and 32 Gbit/s optical data rate [50, 49, 21, 83]. The same command/snoop bus is used in all configura-

Table 6.2: Memory subsystem of the modeled system. In the table, MSHR stands for miss status holding register. Cycle counts are in processor cycles. Bus latencies are contention-less latencies.

Memory Subsystem		
	L2 cache	L3 cache
Caches	16	1
Cache size	2MB	64MB
Cache banks	8	8
Cache associativity	16-way	16-way
Cache access latencies	9 cycles	45 cycles
Cache writeback policy	WB	WB
Cache block size	64B	64B
MSHR entries	64	128
Coherence protocol		MESI
Address-network snoops per cycle		8
Address-network snoop-request latency		8
Address-network snoop-response latency		6
L3/Memory controllers		8
L3/Memory controllers' bandwidth		8x256GB/s
Memory latency		100 cycles

tions and is excluded from the power figures, in order to isolate the contribution of the data networks, which are the subject of our study.

Oblivious, Wavelength-routed network (*Oblivious*): This is the proposed network described throughout Chapter 5. We evaluate three different configurations based on how they use the multiple network layers. All configurations require 16 physical network layers, each embedded with 4 virtual layers, to achieve the target bandwidth. All employ 4 protocol network layers, again with 4 virtual layers. Total of 56 wavelengths are required. In *Oblivious-16*, each message is transmitted over a single network layer, whereas in *Oblivious-8* (*Oblivious-4*) each message transmission uses two (four) layers. Table 6.3 summarizes their main parameters. Note that, path lengths, therefore network latencies, are source-destination dependent. For the layout and routing scheme we

Table 6.3: Evaluated configurations of the proposed network.

Oblivious Data Networks			
Network layers	16x1	8x2	4x4
Virtual layers per network layer	4	4	4
Network bandwidth (TB/s)	6	6	6
First-word transmit cycles[a]	4	2	1
Network latency[b] / Delivery	1-3 cycles / 1 cycle		
Replacement policy	LRU[c]		
Transmitter buffer entries	4		
Protocol network layers	4		
Time-slot duration	4 cycles		
Arbitration cycle (in a time slot)	4. cycle		

[a]Includes E/O delays for first-word bits
[b]Includes 4FO4 + light propagation + O/E delays
[c]16x1 configuration has optimizations described in the text

consider, the average (max) path length is 31.5 mm (67.5 mm). In all configurations, we match the receive bandwidth in a node through four 128-bit delivery ports, each serving a subset of the data network layers. Unless otherwise stated, all optimizations (Section 5.3.5) are employed.

We employ replacement policy optimizations in *Oblivous-16*. Notice that, a memory node can simultaneously accommodate the connections from all sixteen cache nodes on non-conflicting layers. This also eliminates the need for proactive breaks at memory nodes. Similarly, a cache node can simultaneously accommodate the connections from all eight memory controllers on non-conflicting layers; though, they may conflict with connections from other cache nodes. As a result, we use a static node-to-layer mapping in a memory (cache) node to ensure even distribution of cache-to-memory (memory-to-cache) connections across network layers at both sender and receiver side. This results in good load balance and minimal number of connection setups. On cache-to-cache connections, caches uses LRU replacement policy.

In addition, we find that proactively breaking all used cache-to-memory connections (for writebacks) at memory nodes in *Oblivious-4* performs better than proactively breaking only one connection. Writebacks from caches are typically irregular, and reducing the setup delay for several subsequent connection requests outweighs the benefit of otherwise increased connection hit rate, which is already not very high in *Oblivious-4* due to conflicts.

In all other cases, nodes use LRU replacement policy and proactively break only one connection.

Optical crossbar with broadcasting (*Xbar-Bcast*): This optical network is modeled after the data network of Kırman et al. [39]. Its optical fabric essentially implements a full crossbar on a set of waveguides that loop around all nodes. Each source node has exclusive set of wavelengths on which it broadcasts data packets. All nodes other than the sender tap the data, but only the true destination processes it. As a result, the network operation does not require global arbitration. However, this comes at the expense of O(N) number of receivers per node. To mitigate the resulting cost, the authors suggest a hierarchical opto-electrical organization where the optical fabric serves several (electrical) switches at the top level, and each switch serves multiple nodes at the lower level. We perform a similar design-space exploration as in the original work to determine the organization that provides the best power-performance trade-off for our target bandwidth of 6 TB/s (Appendix D). The resulting configuration has 6 switches on the bus, each capable of transmitting 4 messages using 2 wavelengths per message in a waveguide, and a flit size of 64 bytes. Total of 48 wavelengths are used.

Optical crossbar with arbitration (*Xbar-Arb*): This network is modeled after

the data network in Vantrease et al. [72]. It implements a crossbar as well, however this time, each node has an exclusive set of waveguides that loop around all other nodes on which it receives data from the other nodes. The network operation requires arbitration for transmitting to a node which is accomplished through token-based all-optical arbitration. The crossbar comes at the expense of O(N) number of transmitters per node. The target bandwidth is reached with a flit size of 64 bytes satisfied using 64 wavelengths in one data waveguide. We estimate 5-cycle latency for a token to circulate around the nodes for our layout and optical parameters. We assume nodes request one token at a time.

We would like to point out that, because our target system is implemented in an earlier technology node than the one assumed in [72], our results do not necessarily represent the behavior of the system at the scale proposed in that work.

We conduct our evaluation using a cycle-accurate execution-driven simulator based on SESC [56]. Latencies and occupancies of all structures are modeled in detail.

6.2 Applications

We use SPLASH2 applications [76] using MIPS binaries compiled with -O3 optimization level, and data input sets provided in Table 6.4. For the 256-threaded executions, we tried to scale the default data sizes (suggested for up to 64 cores) to account for the four time increase in thread count. We fast-forward the initialization regions (at which point we start modeling timing and collecting statistics) and run them to completion. Our simulation infrastructure currently does

Table 6.4: Applications' simulated problem sizes.

SPLASH-2	Problem size	SPLASH-2	Problem size
Barnes	64k particles	Radix	1,024 radix,
Cholesky	tk29.O		4M integers
FFT	256k points	Raytrace	balls4
LU	1,024×1,024 matrix	Water-NSq	4,096 molecules
Ocean	514×514 ocean	Water-Sp	4,096 molecules

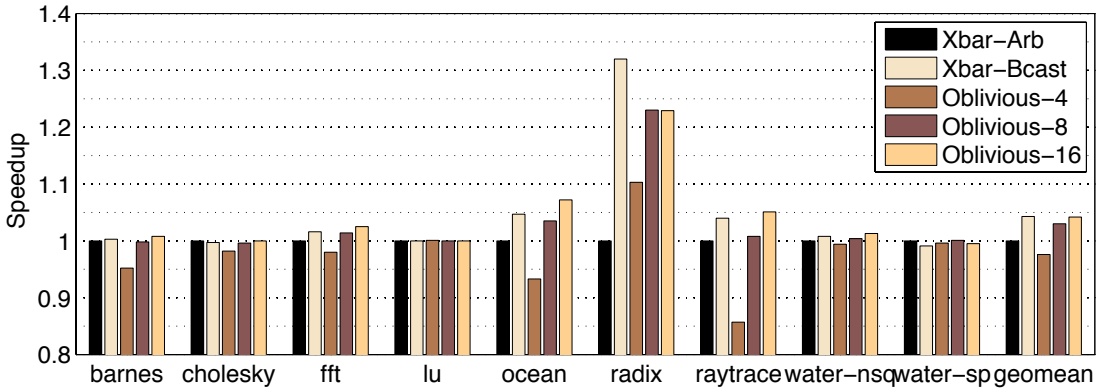


Figure 6.1: Performance of the optical networks relative to Xbar-Arb. In all cases, network bisection bandwidth is 6 TB/s.

not support 256-threaded executions of Volrend and Radiosity. FMM is also excluded due to its long execution time. Already, it is not sensitive to network performance [39, 72].

6.3 Performance Evaluation

Figure 6.1 compares the performance of *Oblivious* to those of *Xbar-Arb* and *Xbar-Bcast*. Speedups are relative to *Xbar-Arb*. Recall that all configurations have the same 6TB/s bisection bandwidth. The results show that all networks are capable to exploit their raw bisection bandwidth to a similar extent.

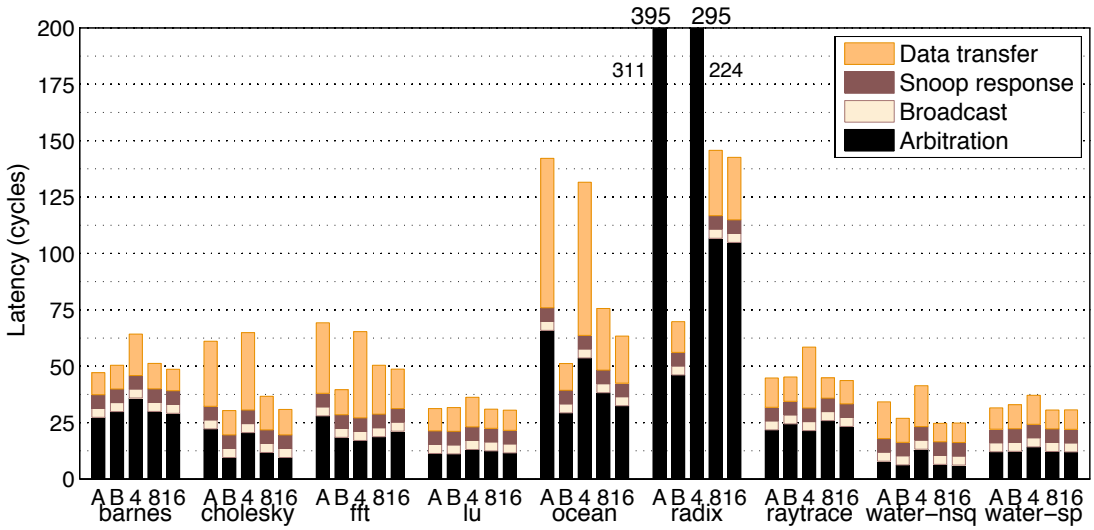


Figure 6.2: Average latency breakdown of a memory operation in the address network and each of the evaluated data networks. Labels A, B, 4, 8, and 16 correspond to Xbar-Arb, Xbar-Bcast, Oblivious-4,-8, and -16 configurations.

We obtain further insights into the performance by examining the average number of cycles a memory operation spends in each data network (Figure 6.2). For reference, we also provide the average latencies for phases on the command/snoop bus. From left to right, the bars for each application correspond to the *Xbar-Arb*, *Xbar-Bcast*, and *Oblivious-4,-8,-16*. We observe very low data-transfer latencies on all networks, *ocean* and *radix* having larger room for improvement.

6.4 Performance Analysis

We conduct additional experiments to gain more insight into the operation of the proposed design.

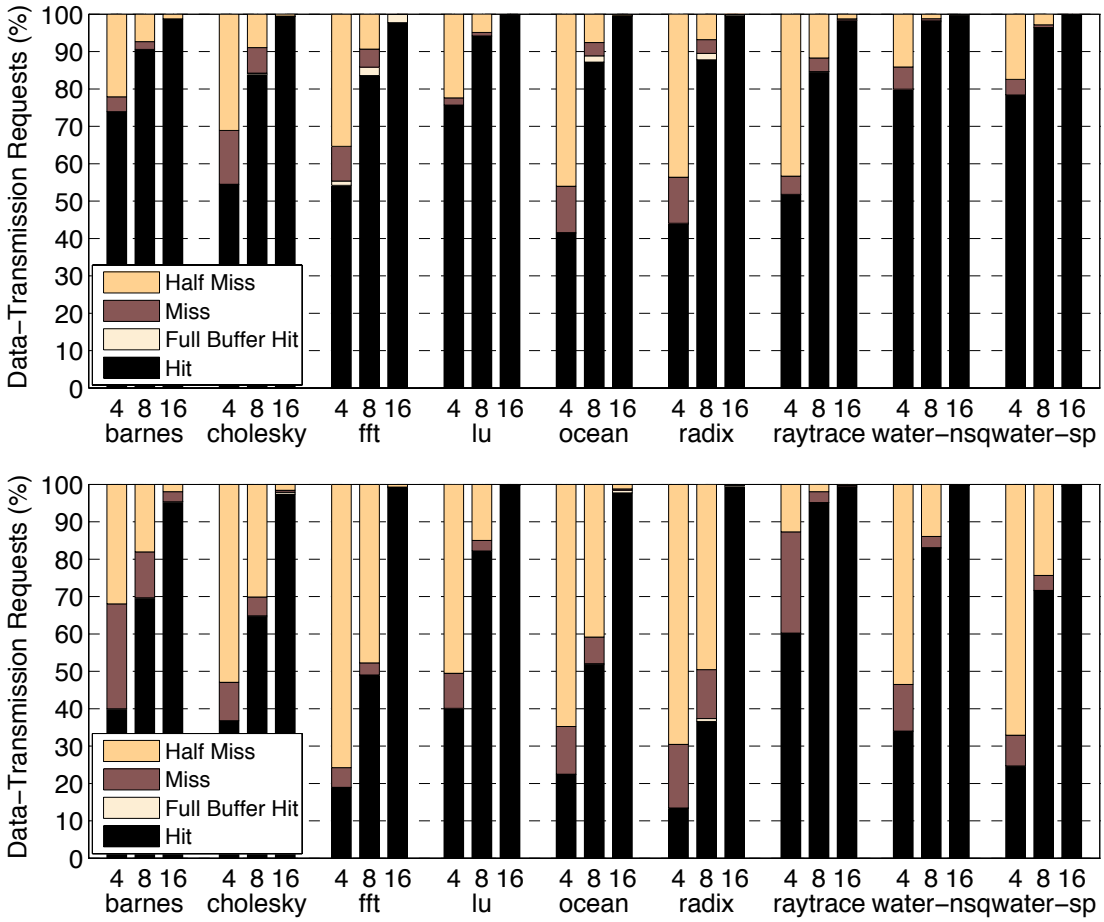


Figure 6.3: Average breakdown of data-transmission requests by a memory controller (top) and by a L2 cache (bottom). The three bars for each application show the results for Oblivious-4,-8,-16, respectively.

The plots in Figure 6.3 break down the true data transmission requests (i.e. excluding lookaheads) by a node based on initially encountered connection state. We show separate plots for memory controllers and L2 caches because of their different characteristics. The three bars for each application show the results for *Oblivious-4,-8,-16* from left to right. A request is classified as *Hit* if the connection exists and there is free space in the transmitter buffer, *FullBuffHit* if the connection exists but there is no buffer space, *Miss* if the connection is un-

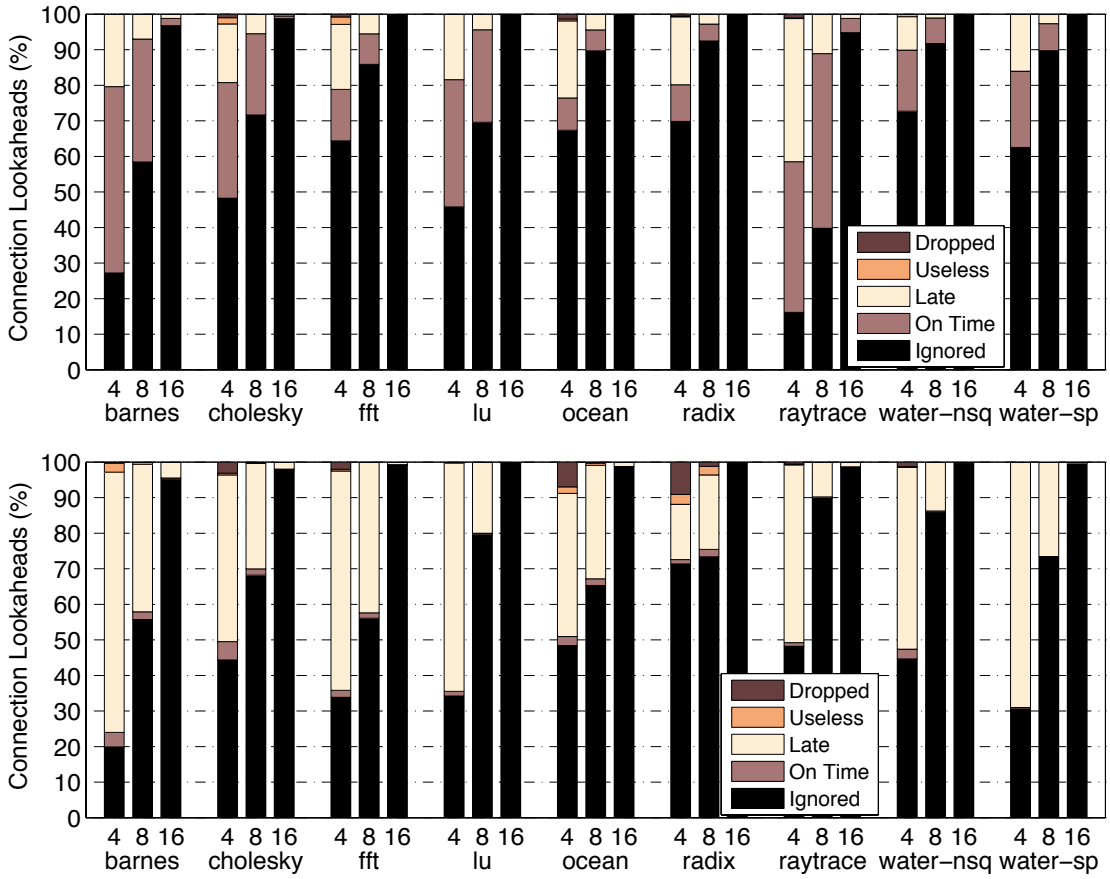


Figure 6.4: Average breakdown of connection-lookahead requests by a memory controller (top) and by a L2 cache (bottom). The three bars for each application show the results for Oblivious-4,-8,-16, respectively.

owned, after which connection establishment is initiated, and lastly *HalfMiss* if the connection is currently being established.

Figure 6.4 has a similar setup as Figure 6.3. The two plots show the breakdown of connection-lookahead requests by a node. *Ignored* encounters valid connection or one currently being established; *OnTime* is successful in setting up a connection before the first use; *Late* establishes a connection but not on time for the first use; *Useless* establishes a connection that is broken before being

used, or is processed later than the true data transmission attempt; and *Dropped* is dropped due to two unsuccessful attempts (Section 5.3.1).

Oblivious-16 simultaneously accommodates all memory-to-cache connections on non-conflicting layers (Section 6.1). These connections may conflict with cache-to-cache connections, which are in minority (less than 35%) except in *raytrace*, *barnes*, and *cholesky*, at receiver cache nodes. As a result, memory nodes have very high hit rates (Figure 6.3, top) and almost all lookaheads are ignored (Figure 6.4, top). *Oblivious-8* owns its high connection hit rates at memory nodes partly to the extensive connection accommodation of the eight layers and partly to the effectiveness of on-time lookaheads (Figure 6.4, top). L3 cache and memory latencies are large enough to hide connection setup delays via lookaheads. The contributions of these two components, however, differ across applications. For example, applications with small amount of cache-to-cache transfers, such as *fft*, *ocean*, *radix*, and *water-**, have high hit rates because a cache node can longer retain the connections from memory nodes on its eight layers. *Oblivious-4* can support less number of connections at a time, resulting in more conflicts and in turn reduced connection hit rates at memory nodes. Lookaheads are still effective.

Turning to the L2 cache-side results, we observe slightly different behaviors. *Oblivious-16* simultaneously accommodates all cache-to-memory connections on non-conflicting layers without other conflicts (Section 6.1). Notice that there is no memory-to-memory data transfers. As a result, *Oblivious-16* can accommodate most of the required connections at a L2 cache at the same time, resulting in near 100% hit rates (Figure 6.3, bottom) and ignored lookaheads (Figure 6.4, bottom). Recall that cache-to-cache connections are typically in mi-

Table 6.5: Average connection statistics, provided separately for L2 caches and memory controllers (L2 cache/Mem Cntr). Connection setup cycles and lifetimes are in processor cycle.

	Oblivious-8			Oblivious-16		
Appl.	**Setup**	**Lifetime (K)**	**Uses**	**Setup**	**Lifetime (K)**	**Uses**
barnes	38/39	3.2/1.7	3/2	35/50	51/46	25/37
cholesky	51/44	3.4/2.9	3/3	36/52	117/137	48/75
fft	50/50	0.8/1.0	2/5	36/47	129/199	132/751
lu	39/38	145/16	6/3	35/43	16187/8305	401/599
ocean	61/51	1.5/0.9	3/7	39/59	84/46	67/221
radix	77/49	0.3/1.5	3/10	34/47	339/1342	732/6821
raytrace	42/39	4.1/0.5	22/2	36/42	76/13	243/18
water-nsq	47/40	4.1/4.9	7/12	34/42	581/223	419/316
water-sp	39/43	2.7/5.3	4/9	35/50	404/701	196/829

nority and although can conflict with memory-to-cache connections at receiver caches, they can be generally accommodated on one of the layers. Our optimizations further help increase connection utilization of cache-to-cache connections, which will be analyzed later. For cache nodes, there is a dramatic reduction in hits for *Oblivious-8* and *-4* due to increased amount of connection conflicts. Gang proactive break at memory nodes in *Oblivious-4* further reduces the hit rates. Most of the missing requests are half misses because the preceding lookahead requests are late (Figure 6.4, bottom). Snoop response and L2 cache read latencies are not long enough to hide the connection setup delay on lookaheads.

Next, we provide connection related statistics in Table 6.5 for *Oblivious-8* and *-16*. We show average connection setup latency, connection lifetime (in Kilo cycles), and number of times a connection is used. The two numbers in each entry correspond to connections established by L2 cache and memory controller, respectively.

In *Oblivious-8*, for example, it takes ~47 cycles on average to establish a con-

Table 6.6: Fraction (%) of all data supplies by sharer caches with existing connection in Oblivious-4, -8, and -16.

Appl.	**Oblivious-4**	**Oblivious-8**	**Oblivious-16**
barnes	55.5	81.1	95.8
cholesky	68.2	85.6	97.7
fft	59.8	81.9	99.2
lu	73.0	91.9	99.7
ocean	65.1	85.1	97.5
radix	50.9	79.2	98.6
raytrace	68.6	96.9	99.6
water-nsq	54.8	90.5	99.4
water-sp	45.8	68.9	97.9

nection. A protocol transaction alone takes ~18 cycles on the time-slotted control network. These correspond to 2.6 transactions per connection establishment on average, demonstrating the effectiveness of proactive breaks. Without proactive breaks, typically a connection setup requires 4 transactions (Section 5.3). Connection lifetimes and connection uses are very high in *Oblivious-16*, as expected, owing to the large connection capacity it can accommodate. For *Oblivious-8*, connections are broken and setup more frequently, resulting in reduced lifetimes and uses.

We next extract the fraction (%) of all data supplies by sharer caches with an already existing connection in Table 6.6. The results clearly show that connection aware shared-data supplier optimization significantly increases connection utilization and improves connection hit rates. Even in *Oblivious-4*, a sharer with an existing connection can be frequently found.

Finally, we run *Oblivious-8* without any optimization, and all combinations where only two of the optimizations are included. Performance improvements in Figure 6.5 are relative to the configuration with no optimization. As a reference, we also provide the original results with all optimizations.

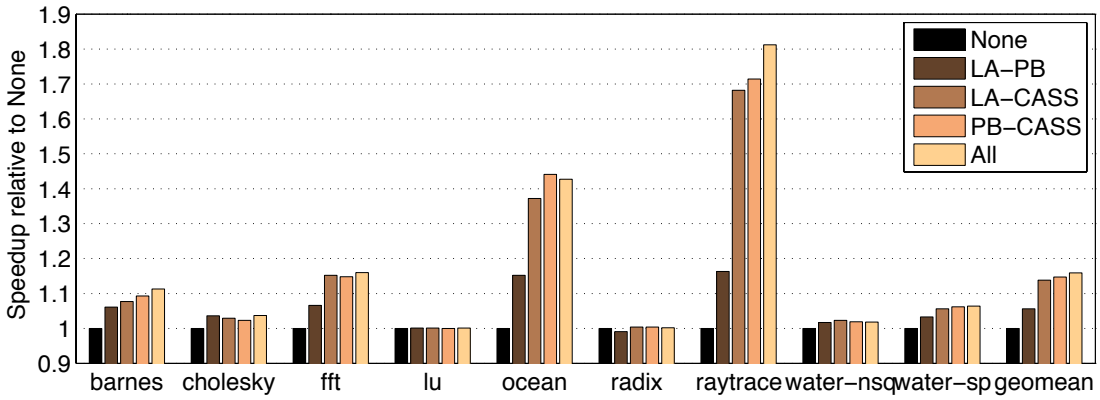


Figure 6.5: Study on effectiveness of connection-aware shared-data supplier (CASS), proactive break (BP), and connection lookahead (LA) optimizations in Oblivious-8.

Connection-aware shared-data supplier (CASS) is most effective, followed by proactive break (PB) support. Overall, we conclude that the three optimizations together are crucial to efficiently exploit the connection-based network operation.

CHAPTER 7

POWER EVALUATION OF OPTICAL NETWORKS*

In this chapter, we estimate the maximum on-chip power consumption of *Xbar-Bcast*, *Xbar-Arb*, and *Oblivious* networks. We also estimate the power for the off-chip laser in each configuration by calculating the optical power requirements. We first describe our methodology and then discuss the results.

7.1 On-chip Electrical Power Estimation

We break down the on-chip power consumption into five categories. Maximum activity factor is assumed (i.e. $\alpha = 1$).

Switches/(De)multiplexers: *Xbar-Bcast* employs electrical routers while all-optical ones (*Xbar-Arb* and *Oblivious*) only have (de)multiplexers at network interfaces. Table 7.1 lists the count, type, and size of these components. Note also that, we account for the data buffers at network interfaces along these structures. We use Orion1.0 [74] to estimate their maximum power consumption.

Wiring: *Xbar-Bcast* consumes additional wiring power on the links from(to) nodes to(from) routers. We estimate wiring power assuming 280 nm global-wire pitch [53] and ITRS device-performance and interconnect projections [36] for power-performance optimized repeatered global wires using the methodology in Ho et al. [33]. Leakage power per repeater is assumed to be 1 μW [39].

Transmitters/Receivers: Following the methodology in [57, 30] and assum-

Table 7.1: Electrical switches/(de)multiplexers in the evaluated networks.

	Electrical Switches/(De)multiplexers	
Xbar-Bcast	6	4x5 routers, 512b, 4-entry input, 1-entry output buffers
	6	21x4 routers, 512b, 4-entry input, 1-entry output buffers
Xbar-Arb	24	1x23 demux, 512b, 1-entry output buffers
	24	1x1 mux, 512b, 2-entry input buffer
Oblivious-16	24	1x16 demux, 512b, 4-entry output buffers
	96	4x1 mux, 128b, 2-entry input buffers
Oblivious-8	24	1x8 demux, 512b, 4-entry output buffers
	96	2x1 mux, 128b, 2-entry input buffers
Oblivious-4	24	1x4 demux, 512b, 4-entry output buffers
	96	1x1 mux, 128b, 2-entry input buffers

ing conservative 100 fF driver plus modulator capacitance and small 2.4 fF photodetector capacitance reported in [21], we estimate 40.5 $\mu W/Gb/s$ and 147 $\mu W/Gb/s$ power at 32 nm technology node for a single transmitter and receiver, respectively. This corresponds to 1.3 mW transmitter power and 4.7 mW receiver power at 32 Gb/s optical data rate. Component counts are provided in Table 7.2. Power estimations consider busy components only. Notice that *Xbar-Bcast* and *Xbar-Arb* have large number of receivers or transmitters. *Oblivious* networks, on the other hand, have just enough components to satisfy the target bandwidth, with a few extra ones in the protocol network layers.

Microrings: Active microrings also consume power. Using the methodology in [46], we estimate dynamic modulation energy to be 82 fJ/bit, assuming V_{on} = 2 V, V_{pp} = 4 V, I_{on} = 50 μA [81, 50], and modulator capacitance of 10 fF [46]. In steady active state, a microring consumes 100 μW power, again assuming these values. This is also in agreement with [64]. Accordingly, we estimate busy ring-resonator-based modulators' and active microring filters' power consumption using the component counts provided in Table 7.2. Passive microrings do not consume electrical power.

Table 7.2: Component counts in the evaluated networks. Counts without parentheses are total component counts, while counts in parentheses show the maximum number of simultaneously active (busy) ones. If only busy component count is provided, it is the total component count as well. Mod. is short for modulators.

	Txs (Busy)	Rxs (Busy)	Microrings	
			Switching (Busy)	Passive
Xbar-Bcast	(1,536)	(7,680)	(1,536) mod.	7,728
Xbar-Arb	35,328 (1,536)	(1,536)	35,328 (1,536) mod. 1,152 (600) filter	1,536
Oblivious-16	(1,920)	(1,920)	(1,920) mod. 11,504 (960) filter	18,879
Oblivious-8	(1,920)	(1,920)	(1,920) mod. 8,816 (768) filter	18,879
Oblivious-4	(1,920)	(1,920)	(1,920) mod. 7,472 (672) filter	18,879

7.2 Optical Power Estimation

Laser sources, which provide light to the on-chip optical network, consume additional electrical power. Among light-source alternatives, we assume off-chip laser(s), which will consume from the system power but not from the CMP's constrained power budget.

Emitted light is first coupled on chip into a power waveguide. Then the power distribution network brings sufficient light to each node on the optical network. Modulated light at nodes is guided to one or multiple detectors depending on the specific network topology. On the way from the laser source to a detector, a light beam encounters various structures such as splitters, merges, bends, crosses, couplers, off- and on-resonance passive or active microrings, modulators, detectors, etc. (e.g. see the light beam in Figure 5.4). In practice,

all such interactions, including the propagation in waveguides, incur losses in the optical power of the light beam. Emitted light power from the laser must be large enough to ensure sufficient optical power reaches detectors after all the power losses on the way.

We perform detailed power-loss analysis for each evaluated optical system and from there determine the required laser power. We compile and use state-of-the-art or projected component efficiencies from recent literature on most common high-index SOI-based silicon photonic technology. We list their corresponding unit losses, in dB, in Table 7.3.

Table 7.3: Loss values used for unit components/events.

Modulator insertion loss (dB)	0.1	[82]
Detector insertion loss (dB)	0.1	[8]
Active ring drop / through / pass losses (dB)	1 / 0.1 / 0.01	[10, 82, 28]
Passive ring drop / through / pass losses (dB)	0.5 / 0.01 / 0.01	[79, 28]
Waveguide propagation loss (dB/mm)	0.1	[15]
90° Waveguide bend loss, 2μm radius (dB)	0.02	[73]
90° Waveguide bend loss, >6.5μm radius (dB)	0.005	[78]
90° Waveguide intersection loss (dB)	0.12	[80]
Waveguide split excess loss / merge loss (dB)	0.04	[51]
Layer-to-layer coupling loss (dB)	1	[29]
Fiber-to-waveguide loss (dB)	0.5	[44]
Laser efficiency (%)		30 [2]
Detector sensitivity (μW)		10 [57, 84, 21]

We first describe the light propagation models for each evaluated network, and then detail the optical power estimation.

Xbar-Bcast Light-path Model

Figure 7.1 shows the light-path model of a 4-node *Xbar-Bcast* network.

In *Xbar-Bcast*, a node broadcasts data on the loop bus on exclusive set of

wavelengths. Therefore, each node fully couples its wavelengths from the power waveguide through passive microrings. The input light to a node is split equally among the multiple waveguides of the loop bus (32 in our case). Modulated light on a waveguide passes from all other nodes, each tapping a -predetermined- fraction of the light. The last of these nodes taps all the light, stopping its circulation. In the figure, only the components on the outermost waveguide are detailed. The figure is also not representative of the actual component counts, but rather reflects the presence and location of these. In our evaluation, we carefully estimate the number of components in each node. We also provide the lengths, merges, and bends of all waveguide segments.

We assume that the power waveguide runs on a separate layer. Optical power is brought to each waveguide on the loop bus through layer to layer coupling. This eliminates the crosses of the loop bus waveguides and the branches of the power waveguide.

Wavelengths of a node have similar paths and therefore power requirements. For each array of microrings encountered on a light path, we conservatively assume the on-resonance microring, if any, to be the furthest one. In reality, microrings for different wavelengths would have a particular order in an array.

Xbar-Arb Light-path Model

Figure 7.2 shows the light-path model of *Xbar-Arb*. The figure depicts a 4-node system for clarity, while we evaluate a 24-node system constructed in the same manner. The figure is not representative of the exact microring counts at nodes, but rather indicates their presence and location.

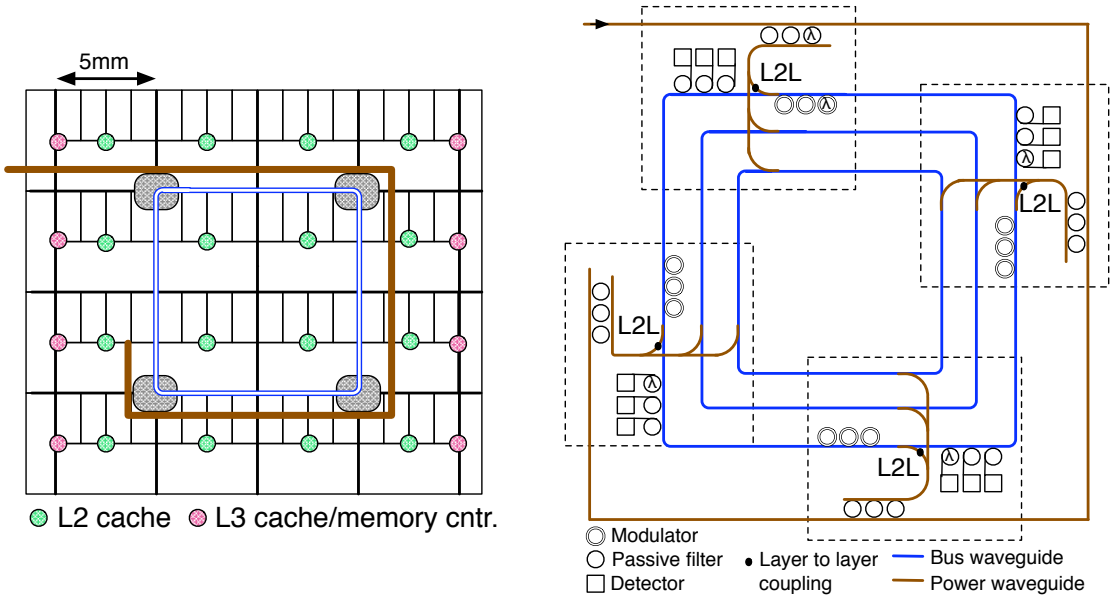


Figure 7.1: Light-path model of Xbar-Bcast.

The architecture comprises a data subnetwork where each node has an exclusive set of waveguides on which it receives data from other nodes. Arbitration for transmitting on a node's waveguides is carried on an all-optical token based arbitration subnetwork.

In the data subnetwork, light beams on all wavelengths are brought to every node through a power waveguide that loops around nodes and branches at each node. The light-power split in a node considers the worst case optical power requirement on the node's waveguides (1 waveguide per node in our case). Light beams propagating in these waveguides are modulated by one of the other nodes and detected back at the home node.

In the arbitration subnetwork, the input light power must be large enough to reach the last node on the power waveguide, be injected as a token on the arbitration waveguide, pass all other nodes, and be absorbed by the same last node. We assume this worst case light path for all token wavelengths. Notice that in a

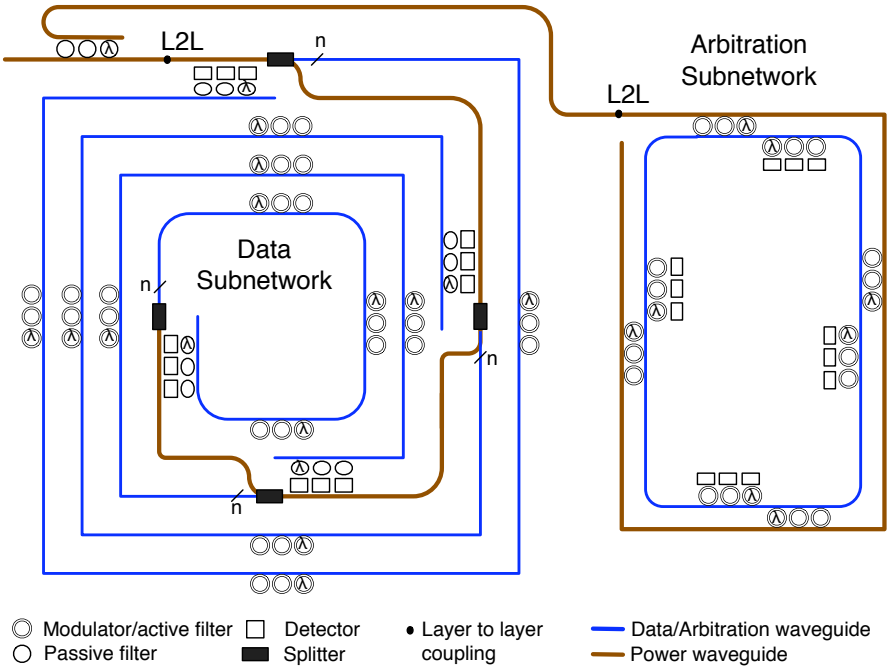


Figure 7.2: Light-path model of Xbar-Arb.

24-node system, the arbitration subnetwork uses only 24 of the 64 wavelengths. In order to optimize optical power consumption, we assume that these 24 wavelengths are extracted from the main power waveguide using passive microrings that couple a predetermined fraction of the on-resonance wavelength. A Y splitter would split all wavelengths with the same splitting ratio.

Xbar-Arb also benefits from layer to layer coupling. It eliminates the crossings between data and power waveguides or arbitration and power waveguides depending on whether the arbitration subnetwork is placed inside or outside of the data subnetwork.

In our 24-node system, power, data, and arbitration waveguides traverse the nodes assuming the same waveguide layout in Figure 5.6. We provide the lengths and bends of all waveguide segments throughout the system, and carefully calculate the number of on- and off- resonance microrings at nodes. No-

tice that the longest light path in both subnetworks loops around the nodes two times. Also, on a data waveguide there is a large number of on- and off-resonance active microrings. We expect these to result in relatively high optical power requirements.

In *Xbar-Arb*, all wavelengths used in a subnetwork have similar paths and therefore power requirements. Again, for each array of microrings encountered on a light path, we conservatively assume the on-resonance microring, if any, to be the furthest one. In reality, microrings for different wavelengths would have a particular order in the array.

Oblivious Light-path Model

A major difference of an *Oblivious* network from *Xbar-Bcast* and *Xbar-Arb* is that, wavelengths are used substantially different in a node. While the light paths of some wavelengths are relatively short or have single destination, some wavelengths are used to communicate to multiple nodes and may have longer paths. This requires wavelength-specific light distribution in order to minimize optical power requirements. We achieve this by first demultiplexing all wavelengths into separate waveguides (Figure 7.3a), which are then routed in parallel over nodes. We assume two power distribution branches, each serving half of the nodes. Demultiplexing the wavelengths onto the waveguides of a power branch is achieved through an array of passive microrings whose power split ratios are set according to the power requirements of the nodes on this branch.

A node splits predetermined and possibly different fraction of light from each of these waveguides (Figure 7.3b). The input light to a node is further

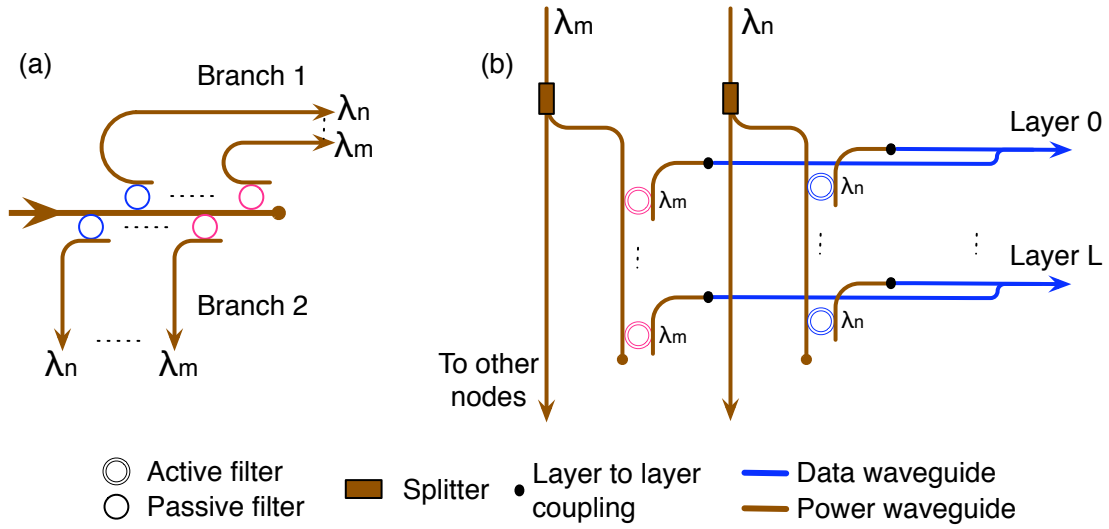


Figure 7.3: (a) Demultiplexing wavelengths into two sets of light-power distribution branches. (b) Distributing light power to individual network layers in a node.

distributed among the input ports to individual network layers. A conservative approach would be to equally split the input power to all layers. Instead, we take a power-aware approach. We observe that (1) a node uses most of the wavelengths to communicate to a single node (Figure 5.1), and (2) communication between a source-destination pair is restricted to one but any of the network layers at a time (Section 5.3). In such a case, only that layer will require the corresponding wavelength. As a result, it is possible to provide the node just enough optical power on a particular wavelength to feed at most as many network layers as the number of total destinations reached by the node on this wavelength. Only the network layers that need a particular wavelength will activate their filter microring on the input power waveguide carrying this wavelength. Other layers will just pass the wavelength. We use binary-tree layout to merge the waveguides for different wavelengths into the single input waveguide to the network layer in order to minimize the merges experienced by a wavelength.

In case multiple network layers are used to transmit a single message, it is enough to split the output light from the filter to these network layers, essentially sharing the filter. Again, we assume binary-tree split layout.

The light injected into a network layer is routed through wavelength routers as described in Section 5.2 and reaches one or multiple detectors.

We provide to the network model the lengths and bend, merge, cross counts for all waveguide segments inside the wavelength routers as well as on the router-to-router and power distribution links. We carefully estimate the count and type (splitting, fully coupling, or passing) of the microrings in every junction at routers by processing the full routing pattern of the network.

A protocol network layer uses only a subset of the wavelengths and light paths based on the time-slot schedule (Section 5.3.3). We estimate the optical power of a protocol network layer excluding the components for the unused wavelengths and light paths.

Similarly as in *Xbar-Bcast* and *Xbar-Arb*, for each array of microrings encountered on a light path, we conservatively assume the on-resonance microring, if any, to be the furthest one. In reality, microrings for different wavelengths would have a particular order in the array. We consider the microring order only when demultiplexing the wavelengths into separate power waveguides.

Power distribution waveguides run on a separate optical layer. Once light is coupled to the drop port of a filter microring in a network layer, it is coupled to the actual network's optical layer. Layer to layer coupling eliminates the crosses between power and data waveguides.

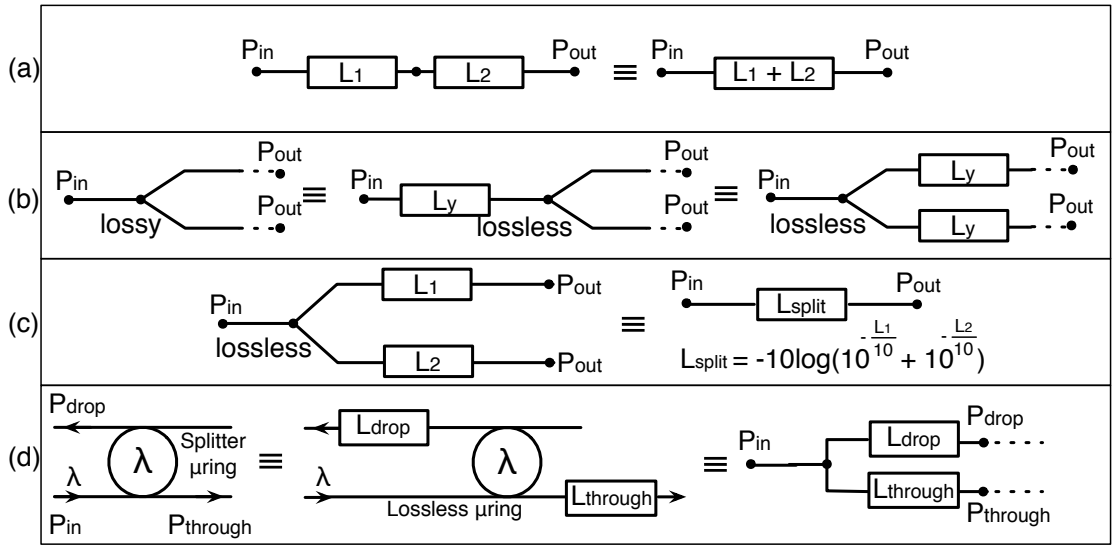


Figure 7.4: Optical loss estimation rules. Losses (L_*) are in dB.

Optical Power Estimation Methodology

We perform detailed power-loss analysis for each network. Starting from the end detectors and walking the light path in reverse direction to that of light propagation, we find the system loss for a particular wavelength up until the off-chip light source by applying the basic loss-estimation rules in Figure 7.4. Using Eq. 7.1, we estimate the corresponding optical power in Watts. In the equation, P_{out} is the optical power required at a detector, and is provided in Table 7.3. L_λ is a negative number as it indicates loss. Then, we sum the optical powers for all wavelength; this is the total optical power the laser(s) need to supply to the system.

$$P_\lambda = P_{out} 10^{-\frac{L_\lambda}{10}} \tag{7.1}$$

Finally, we estimate the electrical power the lasers need to consume to generate sufficient light power. We assume 30% efficiency for a laser [2], which

implies that the lasers will consume ~3.33 times the optical power of the system.

7.3 Results

Following the methodologies described so far, we estimate the on-chip electrical power consumption of all networks as well as the power consumption of the off-chip laser for each configuration. Table 7.4 summarizes these. We report the total optical power the laser(s) need to supply to the system as well as the final laser power consumption.

Table 7.4: Power consumption breakdown for all evaluated networks. Maximum activity factor is assumed (i.e. $\alpha = 1$).

	On-chip Electrical Power Breakdown (W)				Optical Power (W)	Total Power (W)	
	Switches	Wiring	Txs/Rxs	μRings		On-chip	Laser
Xbar-Bcast	39.24	60.40	38.12	4.01	0.91	141.77	3.04
Xbar-Arb	14.37	-	9.22	4.07	90.44	27.66	301.45
Oblivious-16	14.26	-	11.52	5.11	6.13	30.89	20.45
Oblivious-8	8.05	-	11.52	5.09	7.81	24.66	26.03
Oblivious-4	5.01	-	11.52	5.08	8.71	21.61	29.04

The results show that the proposed network is the only one among the evaluated networks that can support very high bandwidths with reasonable electrical and optical power budgets. While *Xbar-Bcast*'s on-chip electrical subcomponents consume a lot of power, *Xbar-Arb* is very sensitive to the efficiency and maturity of the photonic technology.

Xbar-Bcast's power consumption is significantly larger than the one reported in [39]. The reason for this is the very different bandwidth support of the two configurations[1].

[1]We have higher network operation frequency and optical data rate, wider flit width, and

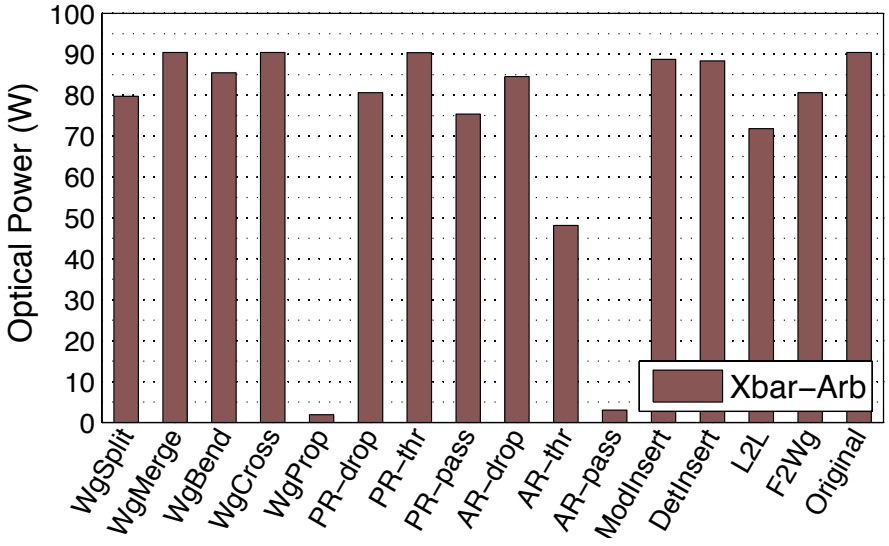


Figure 7.5: Sensitivity of Xbar-Arb's optical power to optical loss parameters. Each category has the corresponding loss parameter set to zero. Wg, PR- and AR-, L2L, F2Wg stand for waveguide, passive ring, active ring, layer to layer coupling, fiber to waveguide coupling, respectively. The optical power with original loss parameters is also provided as a reference (last bar).

We performed a sensitivity study of the optical power to component losses by idealizing (setting to zero) the loss parameters one at a time. Figure 7.5 shows the reductions in optical power for *Xbar-Arb*, while Figure 7.6 shows the results for *Oblivious-8*. *Oblivious-16* and *-4* have similar trends as *Oblivious-8*.

Xbar-Arb is by far most sensitive to waveguide propagation and off-resonance active microring and modulator losses. These are followed by on-resonance active microring and modulator losses. The reason is that, in both data and arbitration parts, the critical paths circulate around nodes two times, once for power distribution and once on actual data or arbitration waveguide. There are many active microrings on the data and arbitration waveguides as well. A possible optimization over the original *Xbar-Arb* layout in [72] is to

larger number of wavelengths.

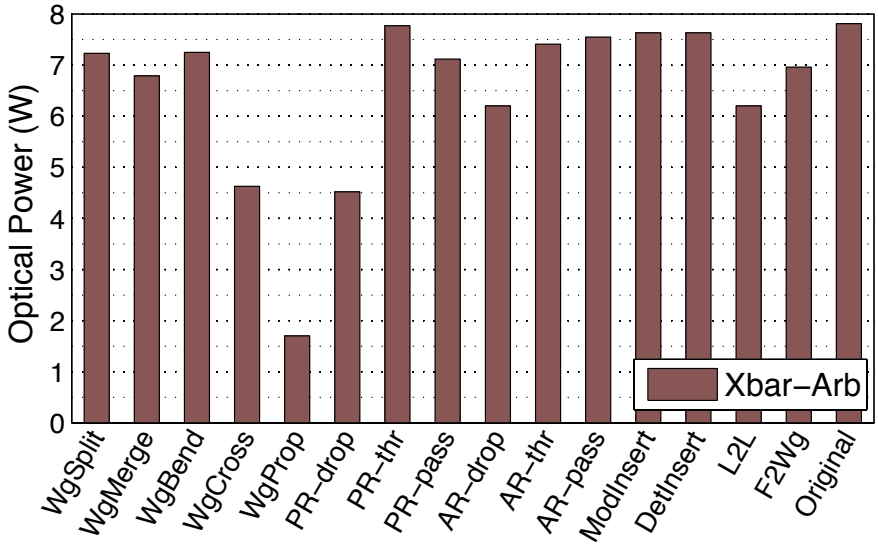


Figure 7.6: Sensitivity of Oblivious-8's optical power to optical loss parameters. Each category has the corresponding loss parameter set to zero. Wg, PR- and AR-, L2L, F2Wg stand for waveguide, passive ring, active ring, layer to layer coupling, fiber to waveguide coupling, respectively. The optical power with original loss parameters is also provided as a reference (last bar).

run the power distribution for the data subnetwork on a second optical layer to shorten path lengths without incurring crossings. When we use two power distribution branches for the data waveguides as in *Oblivious* networks (Figure 5.6) where each branch serves half of the nodes, optical power for *Xbar-Arb* drops to 46.50 W from 90.44 W, and the laser power reduces to 155 W. Still the power requirements are prohibitively high.

The loss parameters that we use are targeted for the most common SOI-based strip waveguides, which allow for very sharp bends–the lowest propagation losses for these waveguides are in the 0.1-0.2 dB/mm range [15]. Ridge waveguides have propagation loss of 0.02 dB/mm. However, they have large pitches and require 200 μm – 600 μm bend-radius [15]. If we would have 0.02 dB/mm

propagation loss, the optical power of *Xbar-Arb* would drop to 3.93 W with the original power distribution layout, and to 3.27 W with the two-branches power distribution (the proposed network's optical power also reduces to 1.77 W, 2.25 W, and 2.5 W for *Oblivious-16*, *-8*, and *-4*, respectively).

Figure 7.6 shows that *Oblivious* networks are most sensitive to waveguide propagation loss, followed by passive microring drop loss, and waveguide intersection loss. The latter two occur mostly in the wavelength routers.

Overall, *Oblivious* offers the best power-performance trade-off among the studied networks, as it has potential to yield significant power savings.

CHAPTER 8

RELATED WORK*

Starting from the recent works, Pan et al. [58] employ optical crossbars in a hierarchical electro-optical topology. Intra-cluster communication is facilitated via an electrical packet-switched network, and inter-cluster communication is carried on multiple optical crossbars, each connecting the routers at the same position of every cluster. The organization retains all of the routers and a lot of the router-to-router wiring of a conventional electrical network, limiting the potential gains that photonics has to offer.

Shacham et al. [62] propose a circuit-switched on-chip photonic network with reconfigurable broadband optical switches. For every data packet, setup and breakdown of an optical path are needed, and these are carried out on an electrical packet-switched network, where each electrical router configures an optical switch. This makes it necessary to transmit data packets of hundreds of bytes on the optical network (well beyond the size of a typical cache block) to amortize the setup/breakdown cost. Control flow is based on a combination of adaptive routing and dropping blocked packets, though the paper does not flesh out the specifics of this mechanism–in particular, how forward progress is guaranteed. We perform power-performance evaluation of this network in Appendix E and indeed observe that this type of operation is not suitable for conventional shared-memory systems where data packets are relatively small.

*

Cianchetti et al. [23] propose another switch-based on-chip photonic network. It uses source-based routing and reconfigurable optical switches to route data. Switch setup is performed by converting the optical control signals that travel along the data to electrical form, and setting up the switch accordingly. Optical data signals must remain steady throughout the control setup (i.e., transmit at the rate dictated by the control network), which may limit effective bandwidth. Contention at output ports is arbitrated, and "losing" packets are electrically buffered if sufficient buffering exists, or outright dropped otherwise. In the face of network-intensive workloads, the network may necessitate large buffering at each switch to reduce packet drop rates and any associated performance loss. Even then, the paper does not flesh out how forward progress is guaranteed in the presence of dropped packets.

Unlike the works above, Vantrease et al. [72] is a fully-optical solution. They propose a high-bandwidth, low-latency optical crossbar that uses token-based optical arbitration to serialize data transmissions to each node. They report significant speedups for SPLASH-2 applications running on a large CMP configuration relative to electrical packet-switched network. Every node has a separate port to all other nodes' data channels, requiring $O(N^2)$ modulators/transmitter, even though only $O(N)$ of them are active at a time. The token-based arbitration can limit effective throughput, especially in light traffic conditions. Also, the large number of components, especially for high node counts, makes the viability of this architecture highly dependent on its ability to rein in the power consumption and signal losses of optical components, which will be heavily dependent on the maturity and efficiency of the optical technology employed.

Haurylau et al. [31] extract the delay, bandwidth density, and power require-

ments that the optical interconnect components must meet in order for on-chip optical interconnects to be comparable with their electrical counterparts. Similarly, Chen et al. [18] project the performance characteristics of future optical devices and then compare the optical and electrical interconnect paths in terms of delay, bandwidth density, and power. They estimate that, for a unit distance at 32nm technology, the delay of an optical interconnect would be approximately 2.2 times faster than an electrical wire. Further they show that, at the same technology node, optical interconnects consume less power, but have lower bandwidth density than their electrical counterparts due to their wider pitches (assuming polymer waveguides).

Kobrinski et al. [40] investigate optical clock distribution and optical global signaling and compare these with their electrical counterparts. They find little power, jitter, or skew improvements from using optics in clock distribution. However, they conclude that by using WDM, optics can be beneficial for global signaling in terms of high bandwidth and low latency. Chen et al. [20] compare four different technologies (electrical, 3D, optical, and RF) for on-chip clock distribution. They also show that because most of the skew and power of clock signaling arise in local clock distribution, there is no significant skew and power advantages of the new technologies, including the optical solution.

Connor [55] reviews the optical interconnect technologies and opto-electronic devices for inter- and intra-chip interconnects, followed by an EDA design flow methodology for optical link designs. The work describes an optical clock distribution network implementation and finds, through circuit simulation, that such a realization can consume significantly less power (five times lower power in case of 64-node H-three at 5GHz) than its electrical counterpart.

The work also proposes a behavioral model of a 4x4 crossbar-like data network, based on wavelength routing that connects four masters to four slaves. However, they do not evaluate its performance in a system.

On-chip transmission-line-based interconnects have also been proposed as alternatives to traditional global wires. These interconnects make use of very wide metal wires so that signals propagate in the high frequency LC domain at near the speed of light [16]. While they do not require any new process to implement, one of their major drawbacks is that they have very low bandwidth due to the large wire width required, which may not be suitable to realize a wide inter-processor interconnect.

There have been many proposals for off-chip optical interconnects targeting shared or distributed memory multiprocessors. We comment on some recent efforts. Louri et al. [47, 48] propose snoopy address sub-interconnects where an optical token circulates around the processors to provide arbitration to transmit the requests through an H-tree like fully optical interconnect. This approach requires modification of the coherence protocol. Webb et al. [75] focus on optical network implementations in large scale distributed shared memory systems. They propose the use of an optical crossbar (implemented using free space optics) for intra-cluster connections, and either crossbar or a point-to-point hypercube optical interconnect that has less connectivity for the inter-cluster connections. Finally, Chen et al. [22], through detailed power models of optical components and network simulation, explore the design space of power-aware opto-electronic off-chip networks. They propose several techniques to dynamically control the power in such networks, achieving significant power savings. Their analysis is performed using both VCSEL-based modulation and off-chip

laser source feeding multiple-quantum-well (MQW) modulators, finding the VCSEL-based solution slightly more power-performance efficient. Note that we assume (on-chip) ring-resonator-based PIN modulators that generally have favorable characteristics over MQW modulators.

Burger and Goodman [14], in an attempt to exploit the high-bandwidth broadcasting capability of optical interconnects (particularly when free-space optics is used), propose a new execution model to reduce serial overheads within a parallel program by having the serial code performed redundantly at any node of a massively parallel multiprocessor/multicomputer system allocated to the program.

Nelson et al. [54] evaluate the performance improvement of replacing global point-to-point electrical wires between the unified front-end and multiple back-ends of a large-scale clustered multithreaded (CMT) processor, where the back-ends are spread across the die, spatially interleaved with caches due to thermal constraints.

Our first work was the first to investigate the design trade-offs, performance benefits, and power and area costs of integrating CMOS-compatible optical interconnect technology into CMPs. Our second work improves on prior works by constructing and evaluating an all-optical on-chip network that uses wavelength-based oblivious routing. It has the potential to support very high bandwidths with relatively low power.

CHAPTER 9

CONCLUSIONS*

In this dissertation, we have investigated the integration of CMOS-compatible photonic technology to construct high-bandwidth, low-latency, and low-power on-chip networks to interconnect the cores and memory modules in future large-scale chip multiprocessors (CMPs).

First, we have devised and evaluated a high-bandwidth and low-latency hierarchical opto-electrical cache-coherent bus whose optical fabric essentially implements full crossbar(s) that do not require global arbitration. It is used for both the command/snoop bus and the data network in the system. By carefully modeling the speed, area, and power characteristics of electrical and recently-developed optical components, and projecting to 32nm technology, we determine that the architecture yields significant performance within reasonable power and area constraints.

We have further improved on the data network. We have taken an all-optical approach to constructing data networks on chip that combines wavelength-based routing, oblivious routing, passive optical routers, and connection-based operation. Our evaluation shows that a careful design based on these features yields a solution that is competitive with prior proposals from the performance standpoint while consuming at least 2.8 times lower power than the competing mechanisms.

---

*

APPENDIX A

CORE FREQUENCY ESTIMATION*

If we set core frequencies based simply on the maximum transistor switching capability projected by ITRS [35], processor frequencies would be unrealistically high (e.g. 22.98GHz at 32nm). Indeed, once we factor in power constraints, feasible frequency levels are much lower. We now extrapolate a trend of future CMP core frequencies that respects such power limitations.

Borkar [12] provides a trend of the leakage power (as a fraction of total power consumption at high temperature) for generations down to 50nm technology (Figure A.1). Using exponential curve fitting, we obtain the value for our 32nm target. These values, however, assume that no leakage reduction technique, which are expected to reduce leakage by 2.5 times or more in future technologies [12], is applied. Consequently, we assume a 2.5 times reduction in leakage for 50nm and smaller feature sizes (Figure A.1).

Using the ITRS-projected maximum total power (189W for 65nm, and 198W for subsequent technologies) and the above leakage power projections, we obtain the peak dynamic power. The consumed dynamic power on a chip can be expressed using a basic formula as follows:

$$P_D = V_{dd}^2 C_g W_g f \sum_i A_i k_i \tag{A.1}$$

where V_{dd} is the power supply voltage, C_g is the total gate capacitance per mi-

*© 2006 IEEE. Reprinted, with permission, from [*International Symposium on Microarchitecture (MICRO)*, Leveraging optical technology in future bus-based chip multiprocessors, N. Kırman, M. Kırman, R.K. Dokania, J.F. Martínez, A.B. Apsel, M.A. Watkins, and D.H. Albonesi, pages 492-503, Orlando, FL, Dec. 2006.]

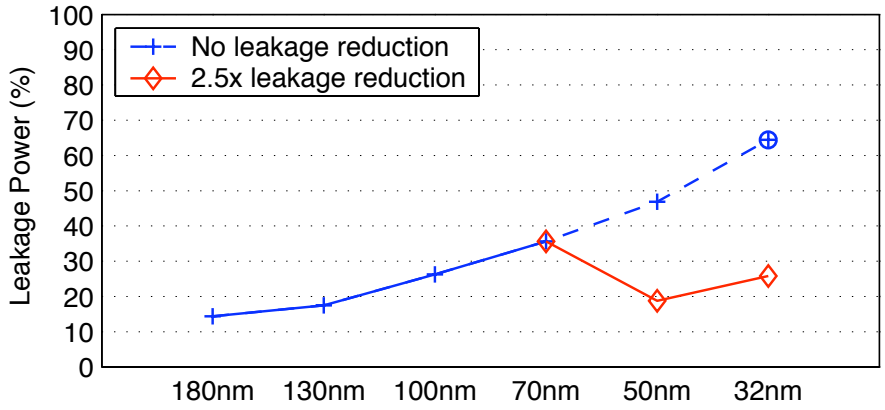


Figure A.1: Leakage power (% of total power) projections, taken from Borkar [12] for up to 50nm technology node, and extended to 32nm using exponential curve fitting. Leakage power percentages assuming 2.5 times leakage reduction for 50nm and 32nm technologies are also plotted.

cron device width ($F/\mu m$), W_g is the minimum transistor width (μm), f is the core frequency, A_i is the switching activity factor for each capacitive circuit node in the processor, and k_i is the ratio of circuit node capacitance to the minimum NMOS transistor gate capacitance, which depends on the circuit topology and transistor sizing, as well as wire capacitance.

We use ITRS projections to set V_{dd} and C_g. W_g decreases by the scaling factor. In the case of $A_i k_i$, we do not use absolute values. Indeed, by assuming that the number of cores, caches, etc. are doubled with each generation while still retaining the circuit structure, we can reasonably assume that the number of circuit nodes also doubles with each generation, thus doubling the sum with each generation. (We also benefit from the fact that local wire capacitance also scales, resulting in the relative ratio of local wire capacitance to minimum gate capacitance to remain constant.)

Table A.1: Summary of ITRS [35] parameters used to calculate the processor frequencies at different technology nodes.

Technology	65nm	45nm	32nm
P_{TOT} **(W)**	189	198	198
C_g **(E − 16F/μm)**	6.99	7.35	6.28
V_{dd} **(V)**	1.1	1	0.9
Frequency (GHz)	4.00	4.40	4.08

Following these trends, substituting the known parameters (Table A.1) in the formula, and assuming a 4GHz core frequency at 65nm, we find that the core frequency remains approximately constant in subsequent technologies (Table A.1, bottom row). This finding is in agreement with Intel's projections in[13]. Thus, in our 32nm CMP model, we assume a processor core frequency of 4GHz.

APPENDIX B

WAVELENGTH PATHS FOUND BY THE GENETIC ALGORITHM

This section provides the full listing of the wavelength paths for the adopted 6x4 torus topology found by the genetic algorithm in Tables B.1 and B.2. Each row corresponds to a transmitter node, each column corresponds to a receiver node. Entry (i, j) gives the <wavelength:path> information for the communication from transmitter i to receiver j. A path is an ordered list of the output ports of the routers visited on the way. E, W, N, S, L corresponds to east, west, north, south, and local port of a router, respectively.

Table B.1: Routing schemes for the 6x4 torus oblivious wavelength-routed optical network found by the genetic algorithm - Part I.

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
N0	0:L	0:EL	0:WWWL	0:EEL	0:WWL	0:WL	0:SL	0:SEL	0:SEEL	0:SEEEL	0:WWSL	13:WSL
N1	1:WL	1:L	1:EEL	1:EL	1:WWWL	1:WWL	1:WSL	1:SL	1:SEL	1:EESL	13:EEESL	0:SWWL
N2	2:WWWL	2:WWL	2:L	2:WL	2:EL	2:EEL	2:SEEEL	2:WWSL	2:SWL	13:SL	1:ESL	0:SEEL
N3	3:WWL	3:WL	3:EL	3:L	3:EEL	3:WWWL	3:WWSL	3:SWL	13:SL	2:SEL	1:SEEL	0:ESEEL
N4	4:EEL	4:EEEL	4:WL	4:WWL	4:L	4:EL	4:EESL	13:WWWSL	3:SWWL	2:SWL	1:SL	0:ESL
N5	5:EL	5:EEL	5:WWL	5:EEEL	5:WL	5:L	13:ESL	4:SEEL	3:SWWWL	2:WSWL	1:SWL	0:SL
N6	6:NL	6:NEL	6:NWWWL	6:EENL	6:NWWL	13:WNL	5:L	4:EL	3:EEL	2:WWWL	1:WWL	0:WL
N7	7:NWL	7:NL	7:ENEL	7:ENL	13:EEENL	6:WWNL	5:WL	4:L	3:EL	2:EEL	1:WWWL	0:WWL
N8	8:NWWL	8:NWL	8:NEL	13:NL	7:ENEL	6:EENEL	5:WWL	4:WL	3:L	2:EL	1:EEL	0:WWWL
N9	9:WWNWL	9:WWNL	13:NL	8:NWL	7:NEL	6:ENEL	5:EEEL	4:WWL	3:WL	2:L	1:EL	0:EEL
N10	10:ENEL	13:WWWNL	9:NWL	8:NWWL	7:NL	6:NEL	5:EEL	4:EEEL	3:WWL	2:WL	1:L	0:EL
N11	13:ENL	10:EENL	9:NWWL	8:ENEEL	7:WNL	6:NL	5:EL	4:EEL	3:WWWL	2:WWL	1:WL	0:L
N12	11:SSL	10:SESL	9:NNWWWL	8:NNEEL	7:SSWWL	6:NWNL	5:NL	4:NEL	3:EENL	2:ENEEL	1:NWWL	12:WNL
N13	11:WSSL	10:SSL	9:SSEEL	8:SESL	7:ENENEL	6:NWWNL	5:WNL	4:NL	3:NEL	2:NEEL	12:EEENL	1:WWNL
N14	11:ESSEEL	10:SWSWL	9:NNL	8:SWSL	7:SSEL	6:SESEL	5:WWWNL	4:NWWL	3:WNL	12:NL	2:ENL	2:EENL
N15	11:WWSSL	10:NNWL	9:SSEL	8:SSL	7:NENEL	6:WNWWNL	5:NWWL	4:WNL	12:NL	3:ENL	3:EENL	3:WWWNL
N16	11:SSEEL	10:ENEENL	9:NNWL	8:NWNWL	7:SSL	6:SSEL	5:NEEL	12:WWWNL	4:WWNL	4:NWL	4:NL	4:ENL
N17	11:NNEL	10:NEENL	9:WWNNL	8:ENNEEL	7:NNWL	6:SSL	12:ENL	5:EENL	5:NWWWL	5:NWWL	5:NWL	5:NL
N18	11:SL	10:ESL	9:WSWWL	8:SEEL	7:SWWL	12:WSL	6:SSL	6:ESSL	6:EENNL	6:NWNWWL	6:NWNWL	6:NWNL
N19	11:WSL	10:SL	9:SEEL	8:ESL	12:EEESL	7:WWSL	7:NWNL	7:NNL	7:ESSL	7:ENENL	7:ENENEL	7:WWSSL
N20	11:WWSL	10:SWL	9:SEL	12:SL	8:EESL	8:EESEL	8:NWNWL	8:NWNL	8:NNL	8:NENL	8:EESSL	8:NWNWWL
N21	11:ESEEL	10:WSWL	12:SL	9:SWL	9:SEL	9:SEEL	9:WNWNWL	9:WNWNL	9:WNNL	9:SSL	9:SESL	9:SEESL
N22	11:SEEL	12:WWWSL	10:WSL	10:WSWL	10:SL	10:ESL	10:ENENL	10:WSWSWL	10:WSWSL	10:WNNL	10:NNL	10:ESSL
N23	12:ESL	11:EESL	11:WWSL	11:WWSWL	11:SWL	11:SL	11:ENNL	11:EENNL	11:WWNWNL	11:WWNNL	11:WNNL	11:SSL

Table B.2: Routing schemes for the 6x4 torus oblivious wavelength-routed optical network found by the genetic algorithm - Part II.

	N12	N13	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23
N0	11:SSL	10:ESSL	9:SEESEL	8:NENEL	7:SSWWL	6:NWNL	5:NL	4:NEL	3:EENL	2:NEEEL	1:NWWL	12:WNL
N1	11:WSSL	10:SSL	9:NEENL	8:SSEL	7:ENENEL	6:NWWNL	5:WNL	4:NL	3:ENL	2:ENEL	12:EEENL	1:WWNL
N2	11:SESEEL	10:SWSWL	9:NNL	8:SWSL	7:SSEL	6:SSEEL	5:WWWNL	4:NWWL	3:WNL	12:NL	2:ENL	2:EENL
N3	11:SWWSL	10:NNWL	9:SSEL	8:SSL	7:NENEL	6:SSEEEL	5:NWWL	4:WNL	12:NL	3:ENL	3:EENL	3:WWWNL
N4	11:SSEEL	10:SWWSWL	9:NWNL	8:NWNWL	7:NNL	6:NENL	5:NEEL	12:WWWNL	4:WWNL	4:NWL	4:NL	4:ENL
N5	11:NNEL	10:NENEL	9:NWWNL	8:NEENEL	7:NNWL	6:SSL	12:ENL	5:EENL	5:NWWWL	5:NWWL	5:NWL	5:NL
N6	11:SL	10:SEL	9:SEEEL	8:ESEL	7:SWWL	12:WSL	6:SSL	6:ESSL	6:EENNL	6:NWNWWL	6:NWNWL	6:NWNL
N7	11:WSL	10:SL	9:SEEL	8:ESL	12:EEESL	7:WWSL	7:NWNL	7:SSL	7:ESSL	7:ENENL	7:WWSSWL	7:WWSSL
N8	11:WWSL	10:SWL	9:SEL	12:SL	8:ESEL	8:EEESL	8:NWNWL	8:NWNL	8:NNL	8:NENL	8:ESESL	8:EEESSL
N9	11:ESEEL	10:WSWL	12:SL	9:SWL	9:ESL	9:ESEL	9:ESESEL	9:WNNWL	9:WNNL	9:SSL	9:ESSL	9:ESESL
N10	11:SEEL	12:WWWSL	10:SWL	10:SWWL	10:SL	10:ESL	10:ENENL	10:SWWSWL	10:SWWSL	10:NWNL	10:SSL	10:ESSL
N11	12:ESL	11:EESL	11:WWSL	11:WWWSL	11:SWL	11:SL	11:ENNL	11:EESSL	11:WWSSWL	11:WWSSL	11:WNNL	11:SSL
N12	0:L	0:EL	0:WWWL	0:EEL	0:WWL	0:WL	0:SL	0:SEL	0:SEEL	0:SEEEL	0:WWSL	13:WSL
N13	1:WL	1:L	1:EEL	1:EL	1:WWWL	1:WWL	1:WSL	1:SL	1:SEL	1:EESL	13:EEESL	0:SWWL
N14	2:WWWL	2:WWL	2:L	2:WL	2:EL	2:EEL	2:ESEEL	2:WWSL	2:WSL	13:SL	1:ESL	0:SEEL
N15	3:WWL	3:WL	3:EL	3:L	3:EEL	3:WWWL	3:WWSL	3:WSL	13:SL	2:ESL	1:SEEL	0:ESEEL
N16	4:EEL	4:EEEL	4:WL	4:WWL	4:L	4:EL	4:EESL	13:WWWSL	3:SWWL	2:WSL	1:SL	0:ESL
N17	5:EL	5:EEL	5:WWL	5:EEEL	5:WL	5:L	13:ESL	4:SEEL	3:WSWWL	2:SWWL	1:SWL	0:SL
N18	6:NL	6:NEL	6:NWWWL	6:EENL	6:NWWL	13:WNL	5:L	4:EL	3:EEL	2:WWWL	1:WWL	0:WL
N19	7:NWL	7:NL	7:ENEL	7:ENL	13:EEENL	6:WWNL	5:WL	4:L	3:EL	2:EEL	1:WWWL	0:WWL
N20	8:NWWL	8:NWL	8:NEL	13:NL	7:ENEL	6:WWWNL	5:WWL	4:WL	3:L	2:EL	1:EEL	0:WWWL
N21	9:EENEL	9:WNWL	13:NL	8:NWL	7:ENL	6:NEEL	5:EEEL	4:WWL	3:WL	2:L	1:EL	0:EEL
N22	10:ENEL	13:WWWNL	9:WNL	8:NWWL	7:NL	6:NEL	5:EEL	4:EEEL	3:WWL	2:WL	1:L	0:EL
N23	13:ENL	10:ENEL	9:WWNL	8:NWWWL	7:NWL	6:NL	5:EL	4:EEL	3:WWWL	2:WWL	1:WL	0:L

APPENDIX C

SUPPORT FOR CREDIT-BASED CONTROL FLOW*

For systems where buffer space at destination is not guaranteed, or delivery rate does not match the receive rate there is need for a control-flow mechanism to ensure that a node schedules data packets for transmission only when there is space in the connection's receiver buffer. Adding credit-based control flow to the proposed optical network can be achieved with little additional support. Each entry in the connection status table is extended with a credit count information. Each receiver also has a record of the owner to whom it will send credits. We leverage the protocol network layers to communicate the credits.

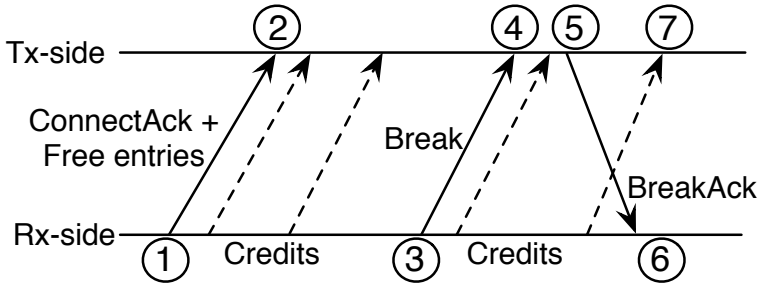


Figure C.1: Credit-flow timing diagram.

For each connection, the credit flow starts with communicating the number of free receiver-buffer entries to the connection owner along with the connection acknowledgement (Figure C.1, (1)). The owner initializes the connection's credit counter (2) which, during the lifetime of the connection, is decremented upon scheduling a data transmission and incremented upon receiving a credit from the receiver. Receiver's node sends back a credit for each data packet it

*

processes and removes from the buffer until it receives either a break acknowledgement or a new connection request from the same owner, whichever arrives first (6). If the connection request arrives first, the owner must have already sent the break acknowledgement for the previous connection. Therefore, it does not need more credits for the previous connection. Notice that, several credits may reach the owner (7) after it sends the break acknowledgement (5). The node does not use these credits; later when it establishes a new connection to the same node, it will initialize the credit counter, ignoring any prior credits. When a new receiver owner is acknowledged and informed of the free entries in the receiver buffer, earliest after receiving the break acknowledgement, the data packets from the previous owner are guaranteed to have occupied entries in the buffer (Section 5.3.1).

Credits are communicated on the protocol network layers. Each outgoing buffer entry is extended with a credit counter, which accumulates the credits while waiting for a proper time slot. These credits are either piggybacked to a protocol message (other than a connection acknowledgement) or sent alone if there is no protocol message. Sending the credits clears the credit count. On transmitting a connection acknowledgement, on the other hand, the credit count is not used but directly cleared; instead, the actual number of free receiver-buffer entries is sent along the message.

APPENDIX D

DESIGN-SPACE EXPLORATION OF XBAR-BCAST OPTICAL NETWORKS

This section provides the results of the design-space exploration to determine the *Xbar-Bcast* organization that provides the best power-performance trade-off for a target bisection bandwidth of 6 TB/s.

We evaluate *Xbar-Bcast* configurations that have different number of nodes on the optical bus. Accordingly the number of wavelengths per node is estimated assuming availability of up to 64 wavelengths. The resulting configurations are *Xbar-Bcast-4N-6D*, *Xbar-Bcast-6N-4D*, *Xbar-Bcast-8N-3D*, *Xbar-Bcast-12N-2D*, *Xbar-Bcast-24N-1D*. A configuration *Xbar-Bcast-nN-mD*, where *n* and *m* each indicates a particular number, has *n* switches on the bus, each capable of broadcasting *m* data messages using 2 wavelengths per message, with a flit size of 64 bytes. As a result, all optical loop buses use total of 48 wavelengths and 32 waveguides. All networks support 6TB/s bisection bandwidth.

Following the methodologies in Section 7, we estimate the maximum power consumption of each *Xbar-Bcast* configuration. Table D.1 lists the count, type, and size of the electrical routers employed in the networks. Transmitter, receiver, and microring counts are provided in Table D.2. Optical power estimations consider configuration-specific topology, waveguide layout, and component counts.

Power consumption of the networks are summarized in Table D.3.

First, we observe that the 24-node *Xbar-Bcast* configuration has prohibitive maximum power consumption, both on chip and by the laser. The exponential

Table D.1: Electrical switches in the evaluated Xbar-Bcast networks.

	Electrical Switches	
Xbar-Bcast-4N-6D	4	6x7 routers, 512b, 4-entry input, 1-entry output buffers
	4	19x6 routers, 512b, 4-entry input, 1-entry output buffers
Xbar-Bcast-6N-4D	6	4x5 routers, 512b, 4-entry input, 1-entry output buffers
	6	21x4 routers, 512b, 4-entry input, 1-entry output buffers
Xbar-Bcast-8N-3D	8	3x4 routers, 512b, 4-entry input, 1-entry output buffers
	8	22x3 routers, 512b, 4-entry input, 1-entry output buffers
Xbar-Bcast-12N-2D	12	2x3 routers, 512b, 4-entry input, 1-entry output buffers
	12	23x2 routers, 512b, 4-entry input, 1-entry output buffers
Xbar-Bcast-24N-1D	24	1x1 routers, 512b, 4-entry input, 1-entry output buffers
	24	23x1 routers, 512b, 4-entry input, 1-entry output buffers

Table D.2: Component counts in the evaluated Xbar-Bcast networks. Counts without parentheses are total component counts, while counts in parentheses show the maximum number of simultaneously active ones. If only busy component count is provided, it is the total component count as well. Mod. is short for modulators.

	Txs	Rxs	Microrings	
	(Busy)	(Busy)	Switching (Busy)	Passive
Xbar-Bcast-4N-6D	(1,536)	(4,608)	(1,536) mod.	4,656
Xbar-Bcast-6N-4D	(1,536)	(7,680)	(1,536) mod.	7,728
Xbar-Bcast-8N-3D	(1,536)	(10,752)	(1,536) mod.	10,800
Xbar-Bcast-12N-2D	(1,536)	(16,896)	(1,536) mod.	16,944
Xbar-Bcast-24N-1D	(1,536)	(35,328)	(1,536) mod.	35,376

increase in the number of receivers on the optical bus is the main reason for the high on-chip electrical power. On the other hand, the optical power requirement is high mainly due to the long waveguide propagation distances: We assume the same data waveguide layout as in *Oblivious* network (Figure 5.6), and that, the power waveguides loop around the nodes as shown in the light propagation model for *Xbar-Bcast* in Figure 7.1. But even if not for the optical power, we do not choose the 24-node *Xbar-Bcast* network because of its high on-chip electrical power consumption.

Table D.3: Power consumption breakdown for the evaluated Xbar-Bcast networks. Maximum activity factor is assumed (i.e. $\alpha = 1$).

	On-chip Electrical Power Breakdown (W)				Optical	Total Power (W)	
	Switches	Wiring	Txs/Rxs	μRings	Power (W)	On-chip	Laser
Xbar-Bcast-4N-6D	38.25	84.55	23.67	4.01	0.38	150.48	1.28
Xbar-Bcast-6N-4D	39.24	60.40	38.12	4.01	0.91	141.77	3.04
Xbar-Bcast-8N-3D	41.31	48.32	52.57	4.01	1.72	146.21	5.73
Xbar-Bcast-12N-2D	46.54	24.16	81.47	4.01	4.58	156.18	15.27
Xbar-Bcast-24N-1D	38.67	-	168.17	4.01	113.21	210.85	377.4

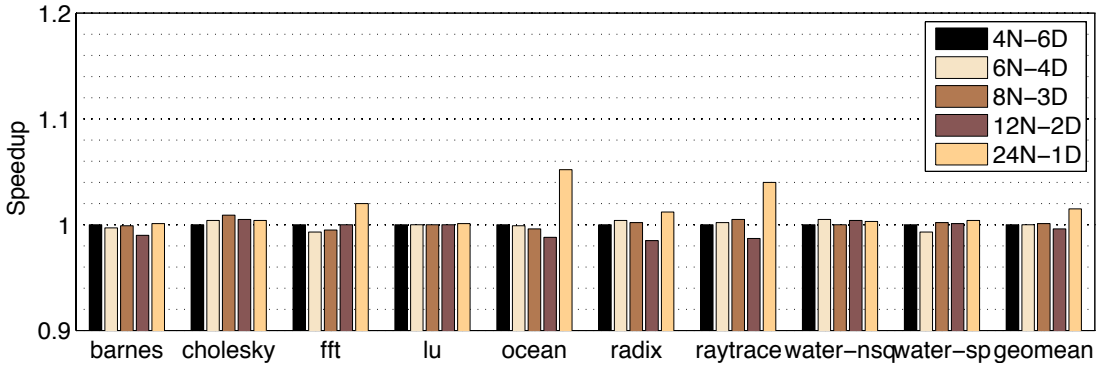


Figure D.1: Performance comparison of the explored Xbar-Bcast networks. Speedups are relative to 4-node configuration, Xbar-Bcast-4N-6D.

The rest of the configurations have similar total power consumption, in the range of 145W-172W. The most power-efficient configuration is the 6-node one, followed by the 4-node and 8-node *Xbar-Bcast* configurations. When going to higher number of nodes on the optical bus, we have less wiring but more receiver components. Also, more optical power is required because of the higher degree of broadcasting and longer propagation distances.

Next, we compare network performances. Figure D.1 plots the speedups of all configurations relative to 4-node *Xbar-Bcast*.

The networks obtain almost the same performance. This is expected, as they have the same operation manner, equal bisection bandwidth, and very small

latency differences. As a result, the 6-node configuration, *Xbar-Bcast-6N-4D*, is the most power-performance efficient configuration among the evaluated ones.

APPENDIX E

EVALUATION OF A CIRCUIT-SWITCHED HYBRID ELECRICAL-OPTICAL NETWORK

We obtain insight into the power and performance of a circuit-switched hybrid electrical network proposed by Shacham et al. [62].

The on-chip photonic network comprises reconfigurable optical switches. The active microrings in a switch are capable of switching all wavelengths at the same time, and are turned ON or OFF based on the desired routing pattern. Optical-path setup and breakdown is needed for every data packet, and it is carried out via an electrical packet-switched network, where each electrical router along the way configures an optical switch.

We configure the photonic network to support the target bisection bandwidth of 6TB/s. The torus network needs to have over-provisioning degree of two in order to achieve the target bisection bandwidth using 64 wavelengths. The layout of the photonic network is shown in Figure E.1. All switches has the same design.

E.1 Performance Evaluation

We evaluate the network's performance by approximating it using our *Oblivious* network configured as follows.

- We evaluate one network layer, as in the original proposal. 64 wavelengths allow the whole cache line of 512 bits to be transmitted in 1 cycle, assuming

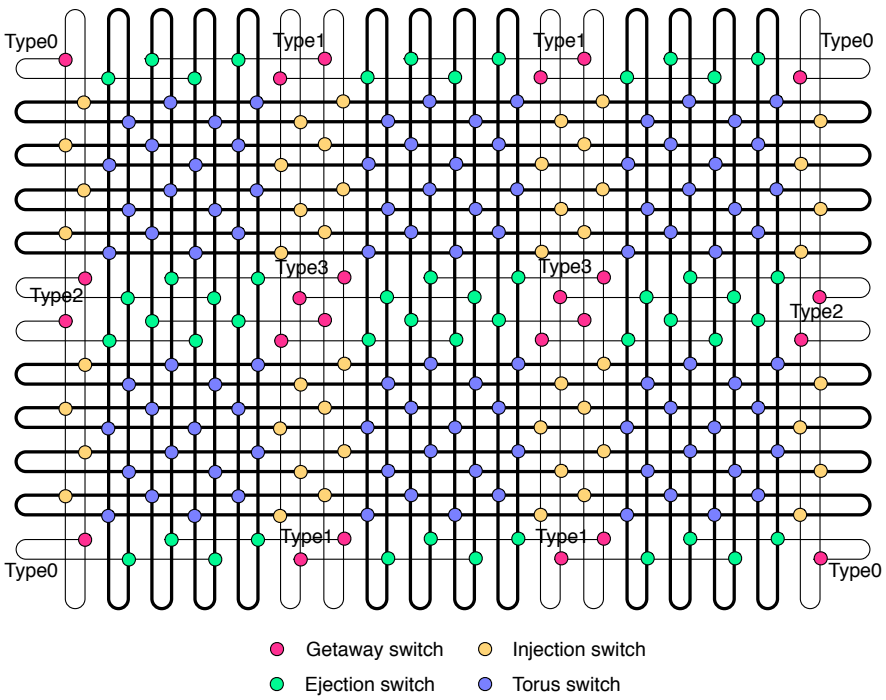


Figure E.1: Layout of the evaluated circuit-switched hybrid electrical-photonic network. It has over-provisioning degree of two. Type0,1,2,3 describe a grouping of nodes based on their longest paths.

32Gbits/s optical data rate. Therefore, we model 1-flit data packets with first-word transmit latency of 1 cycle.

- Network latency between all source-destination pairs is set to 2 cycles. This was estimated based on the propagation delay of light along approximately half of the die perimeter (i.e. 35mm) with enough time left for E/O and O/E conversions.
- The model is changed such that each node handles the transmission of one packet at a time, as in the original proposal.
- We approximate the electrical path-setup network with a constant-delay network that models port contention at nodes and assumes a constant communication latency.

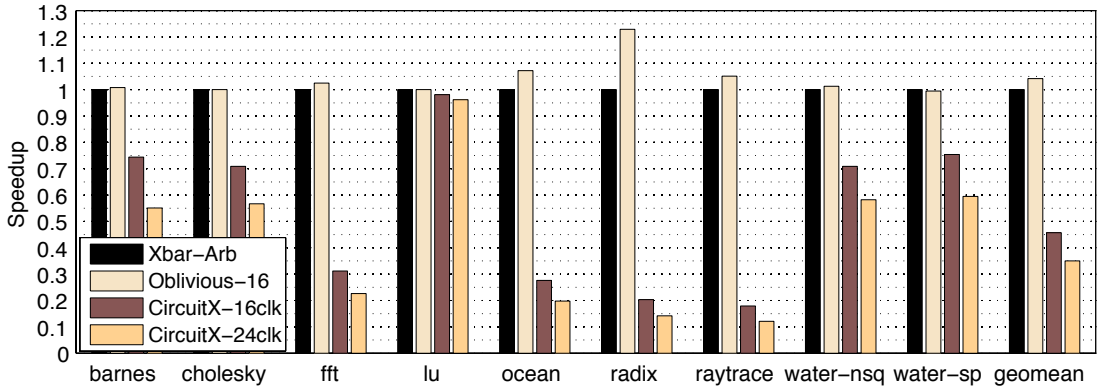


Figure E.2: Approximated performance of the circuit-switched hybrid electrical-photonic network. Speedups are relative to Xbar-Arb.

- Because we use the *Oblivious* network model, the connection-based operation is applied. The original proposal uses a connection for one packet only. To approximate this, we employ proactive break and set the transmitter buffer size to two entries (we set two entries instead of one to compensate for the credit-flow communication delay). When a receiver node looks for proactively breaking a connection, it does not check for the connection to have been used at least once. It is still possible for a connection to be reused multiple times until the connection is broken.

Figure E.2 shows the potential performance of the circuit-switched hybrid electric-photonic network. We evaluate two different configurations of this network, one with 24-cycles and the other with 16-cycles one way path-setup network latency. These correspond to traversal of 6 and 4 electrical-router hops during path set up, assuming 4 cycle hop latency. The speedup is relative to *Xbar-Arb*. The results of *Oblivious-16* are also provided as a reference.

We observe that the circuit-switched network has poor performance. The high photonic-network bandwidth is poorly utilized due to the frequent connection setup delays which are not hidden as well. In addition, there is no parallelism for processing messages in a source node.

We believe that these results are even optimistic. In the original torus network, it is possible to have connection conflicts in intermediate network switches as well. These result in unsuccessful path setups and retrials. Also, the large number of electrical routers in the path-setup network is likely to yield higher hop counts than what we evaluate.

E.2 Power Evaluation

Next, we estimate the maximum power consumption of the circuit-switched hybrid electrical-optical network using the methodology detailed in Chapter 7 and assuming maximum switching activity factor.

The count, type, and size of the electrical routers in the path-setup network are provided in Table E.1. The wiring between routers is also taken into account. Routers are assumed to be connected following the same torus topology as the photonic network. Transmitter, receiver, and microring counts are also provided in Table E.2.

Table E.1: Count, type, and size of the electrical routers in the circuit-switched hybrid electrical-photonic network.

	Electrical Routers
Circuit-X	216 5x5 routers, 32b, 2 VCs, 4-entry input, 1-entry output buffers

We assume the following light-path model in our optical power estima-

Table E.2: Component counts in the circuit-switched hybrid electrical-photonic network. Counts without parentheses are total component counts, while counts in parentheses show the maximum number of simultaneously active ones. If only busy component count is provided, it is the total component count as well. Mod. is short for modulators.

	Txs (Busy)	Rxs (Busy)	Microrings	
			Switching (Busy)	Passive
Circuit-X	(1,536)	(1,536)	(1,536) mod. (1,728) switches	1,536

tions. Light entering the chip is distributed to nodes via two power waveguide branches, each traversing and servicing half of the nodes as in Figure 5.6. The light-power split in a node considers the worst case optical power requirement, which occurs when the node is transmitting to the furthest node. Light beams on all wavelengths are separately modulated by the node's modulators, then routed through the photonic network, and finally filtered out and detected at the destination node. We assume static xy-routing on the torus network as in [62]. Exploiting the symmetry of the network, we observe four different longest paths, and we group the nodes into four categories as well, based on the similarity of their longest paths. We manually extracted the number of crosses, bends, active switch microrings that are passed through and dropped by, and the lengths of each longest path. Then, for each node we used the parameters of the path that is closer to its actual longest path. The layout in Figure E.1 labels the node types. We expect this methodology to closely approximate the optical power that would be found if the actual longest paths had been used. If single longest path was used for all nodes, then the optical power would be much larger than our end result. The network benefits from layer-to-layer coupling by eliminating the crossings between data waveguides and power distribution

waveguides. All wavelengths have the same paths, therefore power requirements. As applied for other networks, for each array of microrings encountered on a light path, we conservatively assume the on-resonance microring, if any, to be the furthest one. In reality, microrings for different wavelengths would have a particular order in the array.

Table E.3: Power consumption breakdown of the circuit-switched hybrid electrical-photonic network. Maximum activity factor is assumed (i.e. $\alpha = 1$).

	On-chip Electrical Power Breakdown (W)				Optical	Total Power (W)	
	Routers	Wiring	Txs/Rxs	μRings	Power (W)	On-chip	Laser
Curcuit-X	13.80	36.33	9.22	4.18	17.25	63.53	57.50

The power breakdown and total power of the network is summarized in Table E.3. The path-setup network contributes approximately 50W of the on-chip power consumption. However, one would expect this sub-network to be lightly loaded because it is unlikely that all routers and links are active since nodes handle one packet at a time. If we assume 0.1 switching activity factor for the path-setup network, routers and wiring would consume 8.53W and 3.98W, respectively, totaling to 12.42W for path-setup network and 25.82W for total on-chip power. This is in the range of *Xbar-Arb* and *Oblivious* networks. The static power consumption in the routers is still significant. We observe that the optical power is also not very low. It is most sensitive to waveguide crossing losses, followed by waveguide propagation losses.

Overall, although the circuit-switched hybrid electrical-optical network has moderate power consumption, its performance is significantly inferior than other optical network alternatives in the context of conventional shared-memory CMP systems.

BIBLIOGRAPHY

[1] A. Aggarwal, A. Bar-Noy, D. Coppersmith, R. Ramaswami, B. Schieber, and M. Sudan. Efficient routing in optical networks. *Journal of ACM*, 43(6):973–1001, November 1996.

[2] J. Ahn, M. Fiorentino, R. G. Beausoleil, N. Binkert, A. Davis, D. Fattal, N. P. Jouppi, M. McLaren, C. M. Santori, R. S. Schreiber, S. M. Spillane, D. Vantrease, and Q. Xu. Devices and architectures for photonic chip-scale integration. *Journal of Applied Physics A: Materials Science & Processing*, 95(4):989–997, June 2009.

[3] V.R. Almeida, C.A. Barrios, R.R. Panepucci, M. Lipson, M.A. Foster, D.G. Ouzounov, and A.L. Gaeta. All-optical switching on a silicon chip. *Optics Letters*, 29(24):2867, December 2004.

[4] H.B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley, Menlo Park, CA, 1990.

[5] K. Banerjee and A. Mehrotra. Power dissipation issues in interconnect performance optimization for sub-180nm designs. In *Symposium on VLSI Circuits Digest of Technical Papers*, pages 12–15, Honolulu, June 2002.

[6] C. A. Barrios, V. R. de Almeida, and M. Lipson. Low-power-consumption short-length and high-modulation-depth silicon electrooptic modulator. *Journal of Lightwave Technology*, 21(4):1089–1098, April 2003.

[7] L. A. Barroso and M. Dubois. The performance of cache-coherent ring-based multiprocessors. In *International Symposium on Computer Architecture*, pages 268–277, San Diego, CA, May 1993.

[8] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kartner, R. Ram, and V. Stojanovic nad K. Asanovic. Building manycore processor-to-DRAM networks with monolithic silicon photonics. In *Hot Interconnects*, pages 21–30, Stanford, CA, August 2008.

[9] A. F. Benner, M. Ignatowski, J. A. Kash, D.M. Kuchta, and M. B. Ritter. Exploitation of optical interconnects in future server architectures. *IBM Journal of Research and Development*, 49(4/5):755, July–September 2005.

[10] A. Biberman, B. G. Lee, K. Bergman, P. Dong, and M. Lipson. Demonstration of all-optical multi-wavelength message routing for silicon photonic networks. In *Optical Fiber Communication Conference*, pages 1–3, February 2008.

[11] M. A. Blake, S. M. German, P. Mak, A. E. Seigler, and G. A. Huben. Bus protocol for a switchless distributed shared memory computer system. United States Patent #6,988,173 B2, International Business Machines Corporation, January 2006.

[12] S. Borkar. Low power design challenges for the decade. In *Conference on Asia South Pacific Design Automation*, pages 293–296, Yokohama, Japan, January–February 2001.

[13] S. Y. Borkar, P. Dubey, K. C. Kahn, D. J. Kuck, H. Mulder, S. S. Pawlowski, and J. R. Rattner. Platform 2015: Intel processor and platform evolution for the next decade. Technical report, Intel White Paper, March 2005.

[14] D. Burger and J. R. Goodman. Exploiting optical interconnects to eliminate serial bottlenecks. In *Proceedings of the Third International Conference on Massively Parallel Processing Using Optical Interconnections*, pages 106–113, October 1996.

[15] J. Cardenas, C.B. Poitras, J.T. Robinson, K. Preston, L. Chen, and M. Lipson. Low loss etchless silicon photonic waveguides. *Optics Express*, 17(6):4752–4757, March 2009.

[16] R. T. Chang, N. Talwalkar, P. Yue, and S. S. Wong. Near speed-of-light signaling over on-chip electrical interconnects. *IEEE Journal of Solid-State Circuits*, 38(5):834–838, May 2003.

[17] A. Charlesworth. The Sun Fireplane system interconnect. In *ACM/IEEE Conference on Supercomputing*, pages 1–14, Denver, CO, November 2001.

[18] G. Chen, H. Chen, M. Haurylau, N. Nelson, D. Albonesi, P. M. Fauchet, and E.G. Friedman. Electrical and optical on-chip interconnects in scaled microprocessors. In *International Symposium on Circuits and Systems*, pages 2514–2517, Kobe, Japan, May 2005.

[19] G. Chen, H. Chen, M. Haurylau, N. Nelson, P. M. Fauchet, E.G. Friedman, and D. Albonesi. Predictions of CMOS compatible on-chip optical interconnect. In *International Workshop on System-Level Interconnect Prediction*, pages 13–20, San Francisco, CA, April 2005.

[20] K.-N. Chen, M. J. Kobrinsky, B. C. Barnett, and R. Reif. Comparisons of conventional, 3-D, optical, and RF interconnects for on-chip clock distribution. *IEEE Transactions on Electron Devices*, 51(2):233–239, February 2004.

[21] L. Chen and M. Lipson. Ultra-low capacitance and high speed germanium photodetectors on silicon. *Optics Express*, 17(10):7901–7906, May 2009.

[22] X. Chen, L.-S. Peh, G.-Y. Wei, and Y.-K. Prucnal. Exploring the design space of power-aware opto-electronic networked systems. In *International Symposium on High-Performance Computer Architecture*, pages 120–131, San Francisco, CA, February 2005.

[23] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi. Phastlane: A rapid transit optical routing network. In *International Symposium on Computer Architecture*, pages 441–450, Austin, TX, June 2009.

[24] J. Crow. Terabus Objectives and Challenges, C2COI Kickoff Meeting, http://www.darpa.mil/mto/c2oi/kick-off/Crow_Terabus.pdf, 2003.

[25] D. E. Culler and J. P. Singh. *Parallel Computer Architecture: A Hardware/Software Approach*. Morgan Kaufmann Publishers, San Francisco, CA, first edition, 1999.

[26] J. D. Davis, J. Laudon, and K. Olukotun. Maximizing CMP throughput with mediocre cores. In *International Conference on Parallel Architectures and Compilation Techniques*, Saint Louis, MO, September 2005.

[27] S. R. Deshpande. Method and apparatus for achieving correct order among bus memory transactions in a physically distributed SMP system. United States Patent #6,779,036, International Business Machines Corporation, August 2004.

[28] D. Ding and D. Z. Pan. OIL: A nano-photonics optical interconnect library for a new photonic networks-on-chip architecture. In *International Workshop on System-Level Interconnect Prediction*, pages 11–18, San Francisco, CA, July 2009.

[29] R. K. Dokania and A. B. Apsel. Analysis of challenges for on-chip optical interconnects. In *Proceedings of Great Lakes Symposium on VLSI*, pages 275–280, Boston, MA, May 2009.

[30] A. Emami-Neyestanak, S. Palermo, H.-C. Lee, and M. Horowitz. CMOS

transceiver with baud rate clock recovery for optical interconnects. In *Symposium on VLSI Circuits Digest of Technical Papers*, pages 410–413, Piscataway, NJ, June 2004.

[31] M. Haurylau, H. Chen, J. Zhang, G. Chen, N.A. Nelson, D.H. Albonesi, E.G. Friedman, and P.M. Fauchet. On-chip optical interconnect roadmap: Challenges and critical directions. In *2nd International Conference on Group IV Photonics*, pages 17–19, Antwerp, Belgium, September 2005.

[32] R. Ho. *On-Chip Wires: Scaling and Efficiency*. Ph.D. dissertation, Dept. of Electrical Engineering, Stanford University, August 2003.

[33] R. Ho, W. Mai, and M. A. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, April 2001.

[34] Intel White Paper. *Next Leap in Microprocessor Architecture: Intel Core Duo*, 2006.

[35] The ITRS Technology Working Groups, http://public.itrs.net. *International Technology Roadmap for Semiconductors (ITRS) 2005 Edition*.

[36] The ITRS Technology Working Groups, http://www.itrs.net. *International Technology Roadmap for Semiconductors (ITRS) 2007 Edition*.

[37] S. Fields J. M. Tendler, S. Dodson. POWER4 system microarchitecture. Technical report, IBM White Paper, October 2001.

[38] P. Kapur, G. Chandra, and K.C. Saraswat. Power estimation in global interconnects and its reduction using a novel repeater optimization methodology. In *IEEE/ACM Design Automation Conference*, pages 461–466, New Orleans, LA, June 2002.

[39] N. Kırman, M. Kırman, R. K. Dokania, J. F. Martínez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi. Leveraging optical technology in future bus-based chip multiprocessors. In *International Symposium on Microarchitecture*, Orlando, FL, December 2006.

[40] M. Kobrinsky, B. Block, J-F. Zheng, B. Barnett, E. Mohammed, M. Reshotko, F. Robertson, S. List, I. Young, and K. Cadien. On-chip optical interconnects. *Intel Technology Journal*, 08(02), May 2004.

[41] R. Kumar, V. Zyuban, and D. M. Tullsen. Interconnections in multi-core

architectures: Understanding mechanisms, overheads and scaling. In *International Symposium on Computer Architecture*, pages 408–419, Madison, Wisconsin, June 2005.

[42] A. F. J. Levi. Fiber-to-the-Processor and Other Challenges for Photonics in Future Systems, http://asia.stanford.edu/events/Spring05/slides/050421-Levi.pdf, 2005.

[43] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, D. Hodge, D. Rubin, U. Keil, and T. Franck. High-speed silicon Mach-Zehnder modulator. *Optics Express*, 13(8):3129–3135, April 2005.

[44] M. Lipson. Guiding, modulating, and emitting light on silicon-challenges and opportunities. *Journal of Lightwave Technology*, 23(12):4222–4238, December 2005.

[45] A. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, and M. Paniccia. A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature*, 427:615–618, February 2004.

[46] J. Liu, M. Beals, A. Pomerene, S. Bernardis, R. Sun, J. Cheng, L.C. Kimerling, and J. Michel. Waveguide-integrated, ultralow-energy GeSi electroabsorption modulators. *Nature Photonics*, 2:433–437, July 2008.

[47] A. Louri and A. K. Kodi. Parallel optical interconnection network for address transactions in large-scale cache coherent symmetric multiprocessors. *IEEE Journal of Selected Topics on Quantum Electronics*, 9(2):667–676, March–April 2003.

[48] A. Louri and A. K. Kodi. An optical interconnection network and a modified snooping protocol for the design of large-scale symmetric multiprocessors (SMPs). *IEEE Transactions on Parallel and Distributed Systems*, 15(12):1093–1104, December 2004.

[49] S. Manipatruni, Q. Xu, and M. Lipson. PINIP based high-speed high-extinction ratio micron-size silicon electro-optic modulator. *Optics Express*, 15(20):13035–13042, October 2007.

[50] S. Manipatruni, Q. Xu, B. Schmidt, J. Shakya, and M. Lipson. High speed carrier injection 18 Gb/s silicon micro-ring electro-optic modulator. In *Proceedings of the IEEE Lasers and Electro-Optics Society*, pages 537–538, Lake Buena Vista, FL, October 2007.

[51] C. Manolatou, S. G. Johnson, S. Fan, P. R. Villeneuve, H. A. Haus, and J.D. Joannopoulos. High-density integrated optics. *Journal of Lightwave Technology*, 17(9):1682–1692, September 1999.

[52] D. A. Miller. Rationale and challenges for optical interconnects to electronic chips. *Proceedings of the IEEE*, 88(6):728–749, June 2000.

[53] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, and ... A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning and 100% Pb-free packaging. In *International Electron Devices Meeting*, pages 247–250, Washington, DC, December 2007.

[54] N. Nelson, G. Briggs, M. Haurylau, G. Chen, H. Chen, D.H. Albonesi, E.G. Friedman, and P.M. Fauchet. Alleviating thermal constraints while maintaining performance via silicon-based on-chip optical interconnects. In *Workshop on Unique Chips and Systems*, Austin, Texas, March 2005.

[55] Ian O'Connor. Optical solutions for system-level interconnect. In *International Workshop on System-Level Interconnect Prediction*, pages 79–88, Paris, France, February 2004.

[56] University of Illinois at Urbana-Champaign. `http://sesc.sourceforge.net`, 2005.

[57] S. Palermo, A. Emami-Neyestanak, and M. Horowitz. A 90nm CMOS 16 Gb/s transceiver for optical interconnects. *IEEE Journal of Solid-State Circuits*, 43(5):1235–1246, May 2008.

[58] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *International Symposium on Computer Architecture*, pages 429–440, Austin, TX, June 2009.

[59] A. M. Pappu and A. B. Apsel. A low power, low delay TIA for on-chip applications. *Conference on Lasers and Electro-Optics*, 1:594–596, May 2005.

[60] P. Rabiei, W. H. Steier, C. Zhang, and L. R. Dalton. Polymer micro-ring filters and modulators. *Journal of Lightwave Technology*, 20(11):1968–1975, November 2002.

[61] A. Rahman and R. Reif. System-level performance evaluation of three-

dimensional integrated circuits. *IEEE Transactions on Very Large Scale Integrated Systems*, 8(6):671–678, December 2000.

[62] A. Shacham, K. Bergman, and L.P. Carloni. Photonic networks-on-chip for future generations of chip multiprocessors. *IEEE Transactions on Computers*, 57(9):1246–1260, September 2008.

[63] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. Davis. Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects. In *IEEE/ACM International Conference on Computer Aided Design*, pages 280–284, San Jose, CA, November 2002.

[64] B.A. Small, B.G. Lee, K. Bergman, Q. Xu, and M. Lipson. Multiple-wavelength integrated photonic networks based on microring resonator devices. *Journal of Optical Networking*, 6(2):112–120, February 2007.

[65] R. A. Soref and B. R. Bennett. Electrooptical effects in silicon. *IEEE Journal on Quantum Electronics*, 23(1):123–129, January 1987.

[66] S.J. Souri, K. Banerjee, A. Mehrotra, and K.C. Saraswat. Multiple Si layer ICs: Motivation, performance analysis, and design implications. In *IEEE/ACM Design Automation Conference*, pages 213–220, Los Angeles, CA, June 2000.

[67] K. Strauss, X. Shen, and J .Torrellas. Flexible snooping: Adaptive forwarding and filtering of snoops in embedded-ring multiprocessors. In *International Symposium on Computer Architecture*, Boston, MA, June 2006.

[68] D. Tarjan, S. Thoziyoor, and N. P. Jouppi. Cacti 4.0. Technical Report HPL-2006-86, HP Laboratories Palo Alto, June 2006.
http://quid.hpl.hp.com:9081/cacti/.

[69] J. Tatum. VCSELs for 10 GB/s optical interconnects. In *IEEE Emerging Technologies Symposium on BroadBand Communications for the Internet Era*, pages 58–61, Richardson, TX, September 2001.

[70] J. M. Tendler, J. S. Dodson, J. S. Fields, H. Le, and B. Sinharoy. POWER4 system microarchitecture. *IBM Journal of Research and Development*, 46(1):5–25, January 2002.

[71] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi. CACTI 5.3,

HP Laboratories Palo Alto,
`http://quid.hpl.hp.com:9081/cacti/`, 2009.

[72] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn. Corona: System implications of emerging nanophotonic technology. In *International Symposium on Computer Architecture*, pages 153–164, Beijing, China, June 2008.

[73] Y.A. Vlasov and S.J. McNab. Losses in single-mode silicon-on-insulator strip waveguides and bends. *Optics Express*, 12(8):1622–1631, April 2004.

[74] H.-S Wang, L.-S. Peh X. Zhu, and S. Malik. Orion: A power-performance simulator for interconnect networks. In *International Symposium on Microarchitecture*, pages 294–305, Istanbul, Turkey, November 2002.

[75] B. Webb and A. Louri. A class of highly scalable optical crossbar-connected interconnection networks (SOCNs) for parallel computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 11(5):444–458, May 2000.

[76] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. The SPLASH-2 programs: Characterization and methodological considerations. In *International Symposium on Computer Architecture*, pages 24–36, Santa Margherita Ligure, Italy, June 1995.

[77] T. K. Woodward and A. V. Krishnamoorthy. 1-Gb/s integrated optical detectors and receivers in commercial CMOS technologies. *IEEE Journal of Selected Topics on Quantum Electronics*, 5(2):146–156, March–April 1999.

[78] F. Xia, L. Sekaric, and Y. Vlasov. Ultracompact optical buffers on a silicon chip. *Nature Photonics*, 1:65–71, January 2007.

[79] S. Xiao, M. H. Khan, H. Shen, and M. Qi. Multiple-channel silicon microresonator based filters for WDM applications. *Optics Express*, 15(12):7489–7498, June 2007.

[80] F. Xu and A. W. Poon. Silicon cross-connect filters using microring resonator coupled multimode-interference-based waveguide crossings. *Optics Express*, 16(12):8649–8657, June 2008.

[81] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson. 12.5 Gbit/s

carrier-injection-based silicon micro-ring silicon modulators. *Optics Express*, 15(2):430, January 2007.

[82] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson. Micrometer-scale silicon electro-optic modulator. *Nature*, 435(19), May 2005.

[83] T. Yin, R. Cohen, M.M. Morse, G. Sarid, Y. Chetrit, D. Rubin, and M. J. Paniccia. 31 Ghz Ge n-i-p waveguide photodetectors on silicon-on-insulator substrate. *Optics Express*, 15(21):13965–13971, October 2007.

[84] T. Yin, R. Cohen, M.M. Morse, G. Sarid, Y. Chetrit, D. Rubin, and M.J. Paniccia. 40 Gb/s Ge-on-SOI waveguide photodetectors by selective Ge growth. In *Optical Fiber Communication Conference*, pages 1–3, February 2008.

[85] T. Yin, A. M. Pappu, and A. B. Apsel. Low-cost, high-efficiency, and high-speed SiGe phototransistors in commercial BiCMOS. *IEEE Photonics Technology Letters*, 18(1):55–57, January 2006.

[86] I. Young. Intel introduces chip-to-chip optical I/O interconnect prototype. *Technology@Intel Magazine*, pages 3–7, April 2004.

[87] H. Zang, J. P. Jue, and B. Mukherjee. A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks. *SPIE Optical Networks Magazine*, 1(1), January 2000.